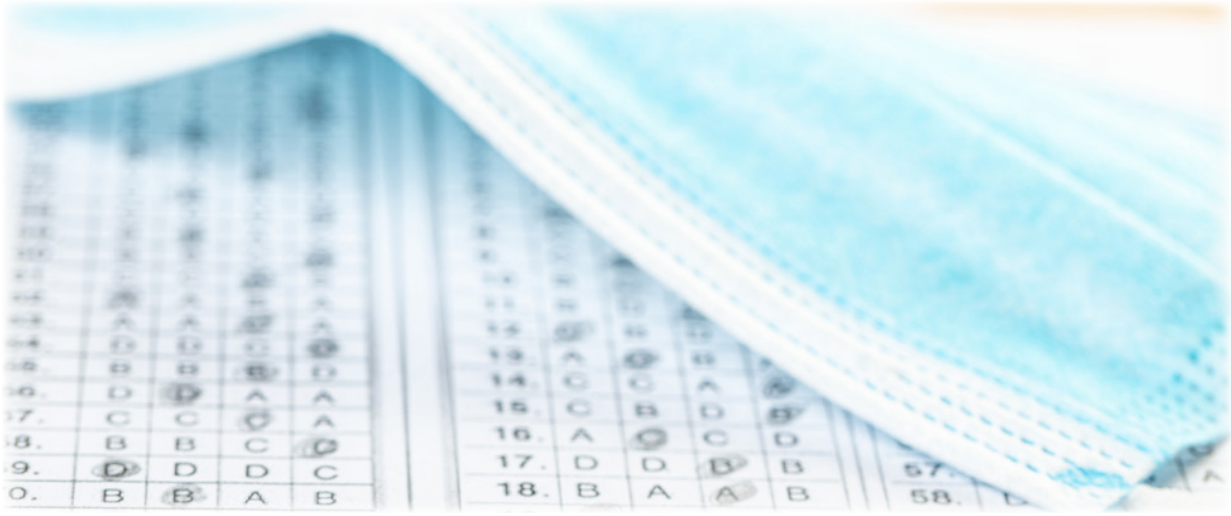
NEPC
NATIONAL EDUCATION POLICY CENTER

Tuesday, March 28, 2023

Newsletter

Five Reasons to Be Cautious When Comparing Pre- and Post-Pandemic Test Scores



As spring state testing season gears up, the results of a new study suggest that educators, policymakers, and parents should exercise caution when using the results to assess the impact of the coronavirus pandemic, or to target resources to student groups with the highest level of need.

The analysis, published in February in the peer-reviewed journal *Educational Measurement: Issues and Practice,* was conducted by NEPC Fellow Benjamin R. Shear of the University of Colorado Boulder. Drawing upon data from math and English/language arts exams administered in 2019 and 2021 to Colorado fifth through eighth graders, Shear explores "the biggest threats to making valid inferences about student learning to study pandemic impacts using state assessment data."

As results from 2023 and beyond become available, the analysis suggests several reasons to be cautious when comparing pre-, during-, and post-pandemic scores.

1. **Test participation rates declined during the pandemic**. For example, in Colorado, the participation rate for the Grade 8 math test declined from 87 percent to 57 percent between 2019 and 2021. "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," Shear writes. "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 prepandemic norms do not represent the full population."

2. **Methods matter**. Shear used three different statistical adjustments to calculate the effect of the pandemic on student test scores. The "Fair Trend" (FT) model compares changes in test scores over time for students with similar assessment results prior to the pandemic, generating the scale scores that students would have been expected to earn in 2021 had they made the same progress their academic peers made prior to the pandemic. "Baseline Student Growth Percentiles" (SGP) use a similar approach, employing percentile ranks to compare "student performance in 2021 relative to their academic peers in a prior 'baseline' cohort of students" who had not yet experienced the pandemic. Finally, a "Multiple Regression" (MR) approach predicts what student test scores would have been absent the pandemic, accounting for pre-existing demographic differences such as socioeconomic status and ethnicity/race. The three methods generate different results, and all three results also differ from simple, unadjusted, year-over-year comparisons. For example, "[r]elative to MR, the FT or SGP approaches will suggest more progress during the pandemic for subgroups that made greater pre-pandemic progress." In addition, Shear notes that "[t]he 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." In other words, simple comparisons of pre- and post-pandemic results may underestimate COVID'S impact.

3. **Changes in scores reflect all experiences that occurred during 2019-20—and not just the impact of the pandemic itself.** For example, during the period of 2011 and 2019, the share of American teens experiencing persistent feelings of sadness of hopelessness rose steadily, from 28 percent to 42 percent. Although the pandemic may have exacerbated this trend, it predated COVID, meaning that changes in test scores between 2019 and 2021 could have been caused—at least in part—by worsening student mental health that was not necessarily associated with the pandemic.

4. **The method(s) selected to calculate the pandemic's impact on test results should depend upon the intended use of the calculation.** For instance, if the goal is to simply target resources to students who made the least progress between 2019 and 2020, the Fair Trend or Student Growth Percentile models may be most appropriate because they treat all students the same rather than accounting for pre-existing demographic differences. However, if the goal is to get an overview of how events that occurred between 2019 and 2020 impacted test performance, it may be more appropriate to use a Multiple Regression model that accounts for pre-existing demographic differences known to have a strong association with test results.

5. **Test results alone cannot gauge the pandemic's impact on "how or why student learning was disrupted, or about impacts on other outcomes such as students' social and emotional well-being."** Other data should also be taken into account, such as information on the type, quality, and duration of pandemic-era remote learning, as well as surveys of student mental health.

NEPC Resources on Assessment

This newsletter is made possible in part by support provided by the Great Lakes Center for Education Research and Practice: http://www.greatlakescenter.org

The National Education Policy Center (NEPC), a university research center housed at the University of Colorado Boulder School of Education, sponsors research, produces policy briefs, and publishes expert third-party reviews of think tank reports. NEPC publications are written in accessible language and are intended for a broad audience that includes academic experts, policymakers, the media, and the general public. Our mission is to provide high-quality information in support of democratic deliberation about education policy. We are guided by the belief that the democratic governance of public education is strengthened when policies are based on sound evidence and support a multiracial society that is inclusive, kind, and just. Visit us at: http://nepc.colorado.edu