Comparing Test Score Growth Measures Constructed from Aggregate and Individual Data

Sean F. Reardon
Stanford University

John Papay
Brown University

Katharine Strunk
Joshua Cowen
Tara Kilbride
Michigan State University

Lily An
Kate Donohue
Brown University

1. Introduction

One of the features of the modern US educational system is the extent to which students' skills are assessed using standardized test scores. State standards-based reform efforts and the federal No Child Left Behind (NCLB) Act led to widespread interest in how to use such test scores to hold schools accountable. Over the past decade, a consensus has emerged among researchers and policymakers that measuring the level of student academic proficiency in a given school is a poor proxy for school quality because average test scores reflect not only the inputs of a school, but also out-of-school factors that shape students' opportunities to learn. Thus, policymakers have begun relying more heavily on student growth, seeking to measure the effectiveness and quality of a school by assessing how quickly its students are learning new material.

Ideally, we would measure the test-score growth for all students in a school in a given grade and year. The presence of out-of-school factors affecting student learning implies that a test-score growth measure would not provide an unbiased estimate of the *causal* effect of the school itself, but it could provide a measure of how much students were learning over the course of the year.

Unfortunately we can only rarely observe current and prior year test scores for all students in a school; student mobility across schools, districts, and states complicates efforts to develop student growth estimates. Given such limitations, there are several ways to operationalize a measure of average learning rates. The best possible feasible approach is to compare end-of-year scores from a given year to scores from the previous year for students who took both tests. Averaging this measure over multiple grade-years provides a *longitudinal growth measure* of the average rate at which students in a given school learn the tested material.

Constructing such measures, however, requires longitudinal student data, which are often not readily available. It is easier to obtain the average student test scores within a school-grade-year. From such data we can compute a different measure: the difference between average scores of all students in a specific grade in a school and the average scores of students in the previous grade in the prior year. This estimate provides a *cohort growth measure* that documents how much student test scores changed, on average, from 3rd grade in one year to 4th grade in the following year, for example. If the exact same sets of students are tested in both years in sequential grades, the longitudinal and cohort growth estimates are exactly the same. However, if some students were tested in one year but not the other, the difference in average scores obtained from the cohort growth approach may not match the average gain score provided by the longitudinal measure.

When longitudinal data are not available, the cohort growth estimate may be the only feasible approach. Indeed, this is the approach used in the Stanford Education Data Archive (SEDA). SEDA is based on the ED*Facts* data, which include aggregate test score data from virtually every public elementary and middle school in the US from 2008-09 to 2016-17. Aggregated scores are available at the school-grade-year-subject-subgroup level and represent over 300 million individual test scores.

The SEDA cohort growth estimates, if valid, may be particularly valuable as they enable comparisons about test score levels and cohort growth across states; in contrast, estimates of longitudinal growth are only available in certain states and any inferences are only valid in comparison to other schools in the same state (Fahle et al., 2018). Thus, understanding how much researchers and policymakers can rely on

these cohort growth estimates requires us to know how different cohort growth estimates are from longitudinal growth estimates and under what conditions they align well.

We address these questions by using longitudinal student data from Massachusetts, Michigan, and Tennessee to construct both longitudinal and cohort growth estimates and assess how well the latter replicates the former. We do this separately for districts and schools. We assess how similar these estimates are in two ways: we estimate the correlation between the two and their root mean square difference.

Intuitively, we expect the two growth measures to align well so long as the groups of students in each cohort do not change much each year. However, the two measures may differ in schools and districts with higher mobility rates or larger gaps in performance across mobile and non-mobile students. Because there is generally more mobility in and out of schools than districts, we might expect that the measures will align better for districts than schools. The effects of mobility may compound over grades, so schools that span more grade levels may be particularly susceptible to large discrepancies between cohort and longitudinal growth estimates. We test these hypotheses.

Findings

The correlations between longitudinal and cohort growth measures are generally strong. On average, SEDA-style cohort growth measures largely rank schools and districts consistently with longitudinal growth measures. For most districts, the discrepancy between the two types of estimates are very small, suggesting that cohort growth is a good proxy for longitudinal growth. However, for about a quarter of districts, the discrepancy is large enough to warrant concern. Correlations are somewhat larger for districts (r=0.87) than for schools (r=0.80), and the root mean squared difference between the two estimates is smaller, on average, for districts than schools.

Furthermore, as expected, the correlations in districts and schools with higher student mobility are somewhat weaker than in those with lower mobility (r=0.84 for districts and r=0.75 for schools with more than 15% annual mobility).

Thus, we conclude that: a) on average, cohort growth measures are useful proxies for longitudinal growth measures; and b) the cohort measures provide useful estimates of longitudinal growth in all but the smallest schools and districts or in schools with a grade span of more than four tested grades.