AERA Open
July-September 2020, Vol. 6, No. 3, pp. 1–4
DOI: 10.1177/2332858420939848
Article reuse guidelines: sagepub.com/journals-permissions

What Is the Stanford Education Data Archive Teaching Us About National Educational Achievement?

Andrew D. Ho
Harvard Graduate School of Education

The Stanford Education Data Archive (SEDA) launched in 2016 to provide nationally comparable, publicly available test score data for U.S. public school districts. I introduce a special collection of six articles that each use SEDA to lend their questions and findings a national scope. Together, these articles demonstrate a range of uses of SEDA for educational research. I review their contributions and discuss next steps for SEDA-based research as more years and levels of data become available.

Keywords: *achievement, educational policy, NAEP, psychometrics, secondary data analysis, testing*

Public school students in Grades 3 to 8 take roughly 45 million math and reading tests each spring. Their scores on these tests are potentially useful measures of aggregate educational opportunity and achievement. However, their usefulness has historically been constrained by the lack of comparability of state test scores across states, grades, and years, and by the fact that states coarsen publicly reported scores into ordered categories (e.g., "below basic," "basic," "proficient," and "advanced") that are neither transparent nor comparable across tests.

The Stanford Education Data Archive (SEDA; Reardon, Ho, et al., 2019) remedies these issues by making average test scores available to the public, by school, district, and county, and by grade, year, subject, and subgroup. There are currently 8 years of district-level data available (spring, 2009 through spring, 2016) in six grades (3–8). Two more years of data will be available later this year. These average scores were estimated using statistical methods for coarsened data (Ho & Reardon, 2012; Reardon et al., 2017) and linking methods for aggregate data (Reardon, Kalogrides, & Ho, 2019). These methods enable estimation of district test score means and standard deviations on the common scale of the National Assessment of Educational Progress (NAEP). The data can be explored and downloaded at https://edopportunity.org.

SEDA data fill a research void at the district and county levels nationally, between coarse state-level research that typically uses NAEP data, and finer, but typically narrower analyses using data systems from one or a few states. This is an important void to fill, given that districts vary considerably in sociodemographic characteristics and have considerable control over education policy (Whitehurst et al., 2013). SEDA also harnesses the stability of the NAEP scale over time, allowing for estimation of progress within states whose testing programs have changed from year to year. These changes have been common under the volatile context of state testing policy this past decade.

The six articles that comprise this special collection showcase the benefits and flexibility of SEDA's national scope and multiyear, multilevel perspective. The range of questions is considerable, as authors explore the relationships between educational opportunity and, respectively: school funding, immigration enforcement, school discipline, parental socioeconomic status (SES), and the choice of measurement instruments themselves.

County-Level SEDA Analyses Using Event Timing

District and county identifiers in SEDA enable researchers to merge test score data to other data sets to answer a broad range of questions. Two articles in this collection accomplished this with county-level data using an event-timing approach.

Shores and Steinberg (2019) study national educational achievement in the wake of the Great Recession by merging county-level SEDA data with the Quarterly Census of Employment and Wages. They define four levels of recession intensity for counties by the relative change in the number of employed workers from the prerecessionary period from spring 2003–2006 to the postrecessionary period spring 2007–2010. They ask whether years of school-age exposure to the 2-year period of recessionary shock, from academic years 2007–2008 to 2008–2009, where school funding cuts would have a differential effect, is associated with lower

educational outcomes in subsequent years. Shores and Steinberg conclude that those exposed to the most severe recession intensity had substantially lower scores, on the order of 0.10 standard deviation units.

The SEDA data window that begins in the academic year 2008–2009 prevents Shores and Steinberg (2019) from implementing a more straightforward differences-in-differences approach that would use prerecessionary trends to target causal inferences. The authors are appropriately descriptive rather than prescriptive about the patterns they observe, particularly when it comes to school funding. Their patterns are nonetheless convergent with causal estimates of changes in per pupil spending on educational test scores, like those of Lafortune et al. (2018) and Jackson et al. (2018).

Bellows (2019) merges in county-level data about the "Secure Communities" immigration enforcement policy to explore whether educational achievement was lower in years following policy activation. She shows that activation was staggered across counties from 2008 to 2013, and she uses this staggered rollout in a difference-in-differences design to identify the relationship between activation and achievement. She finds that activation of Secure Communities is associated with a small reduction in average achievement for Hispanic and non-Hispanic Black students in English language arts (ELA) of around 0.01 standard deviation units, with no statistically significant difference in math. She cautions that prior achievement and enrollment trends, as well as the confounding of activation with other county characteristics, prevent strong claims about the causal effects of policy activation.

District-Level SEDA Analyses Using Adjusted Correlations

A flexible, open data set like SEDA enables different research teams to take various approaches to answering similar questions using the same data. When answers differ, the common data set helps to explain these different answers as functions of differences in questions and approach. Like Bellows (2019), Kirksey et al. (2020) also contribute findings about immigration enforcement and achievement to this special collection. They find that district-proximal enforcement predicts White–Hispanic gaps that are larger in math but not ELA, an apparent contrast with the Bellows finding.

Kirksey et al. (2020) observe correctly that an essential contrast with Bellows (2019) is their different definitions of immigration enforcement. Bellows uses a county-level dichotomous predictor for program rollout, whereas Kirksey et al. use both the number of deportations and the proximity of districts to the site of deportations. They predict that two districts that differ by 850 deportations (1 standard deviation in their data) within 25 miles of the school districts, White–Hispanic ELA gaps differ by 0.28 standard deviation units.

The two approaches have interesting tradeoffs. If their conclusions could have been causal, Bellows (2019) would describe the effects county-level policy activation, and Kirksey et al. (2020) would describe the effects of district-level enforcement. I think Bellows comes closer to a plausibly causal estimate, whereas Kirksey et al. provide a more descriptive finding closer to plausible mechanisms of deportation-related stress and trauma. Neither article provides a compelling reason for the differences in statistical significance across academic subjects ELA and math. To the extent that both ELA and math are proxy measures for educational opportunity, the subject inconsistencies both within each article and between their two articles raise concerns that these findings are within margins of measurement error.

Pearman et al. (2019) take a similar approach to Kirksey et al. (2020), as both articles merge SEDA's educational gaps with other district-level data sets and adjust district relationships for covariates. Pearman et al. merge SEDA gaps with discipline gaps from the Civil Rights Data Collection in the academic years ending 2012 and 2014. They define discipline gaps as the difference in suspension rates, and they leave relative risk ratios (the ratio of suspension rates) as a specification check. They find district-level correlations between educational gaps and discipline gaps of .25 and .29, for Black–White and Hispanic–White gaps, respectively. They also show that district demographic and community characteristics fully predict associations between education gaps and discipline gaps for Hispanic and White students but not Black and White students. In fully adjusted models, they find that a district difference of 1 percentage point in the Black–White suspension rates predicts a district difference of 0.01 standard deviations in Black–White test scores.

Multilevel SEDA Analyses

Jang and Reardon (2019) use the district-level SES measures from the American Community Survey that are native in SEDA to explore district-level relationships between SES and achievement by state. Though these relationships are positive in all states, the states with the strongest relationships have gradients three times steeper than states with the weakest relationships. Jang and Reardon also show that relationships change across grades in varied ways across states. States like Delaware, Vermont, Florida, New York, and New Hampshire have gradients in Grade 8 that are twice the gradients of their respective Grade 3 baselines. This implies that average student learning rates from Grades 3 to 8 are much more strongly correlated with district SES in these states than in others.

Kuhfeld et al. (2019) also use multilevel models to compare SEDA estimates with district-level achievement estimates from another measure: an interim assessment known

as "MAP Growth" from the assessment organization NWEA. MAP Growth is administered in fewer districts nationwide, between 6% and 18% of SEDA districts depending on the year and grade. However, the MAP Growth test has a common score scale that enables comparisons across states, grades, and years, with different assumptions than the cross-test linking procedures described by Reardon, Kalogrides, et al. (2019). They take particular interest in inferences about district growth across grades, given that SEDA data uses a linearly interpolated NAEP scale between Grades 4 and 8, whereas MAP Growth scales have for a stronger basis for grade-to-grade comparisons.

Kuhfeld et al. (2019) fit a version of the SEDA pooling model to MAP Growth data to compare district estimates. They find very strong correlations between SEDA and MAP Growth data for levels of district achievement, .98 and .97 for Math and ELA, respectively. They also find strong correlations for growth estimates, .90 and .82 for Math and ELA, respectively. These correlations are similar to the .87 correlation that Reardon, Papay, et al. (2019) found between district-level growth estimates for SEDA and growth estimates from state longitudinal data systems.

Notably, Jang and Reardon (2019) find interesting patterns in the changes in SES-achievement gradients across grades. For example, the Grade 3 gradient is negatively correlated with the rate of change of the gradient from Grade 3 to Grade 8 ($r \approx -.19$). This means that states with flatter (more equitable) SES-achievement gradients in third grade, on average, increase their gradients. Because Kuhfeld et al. (2019) find that their estimates differ from SEDA growth estimates in high socioeconomic status districts and in linearity in higher grades, I would be interested to see their replication of Jang and Reardon's state-level correlations, if MAP Growth data are sufficiently representative in enough states.

Future SEDA Analyses

I notice three common themes among these articles that indicate the continued promise of SEDA and similar datasets that are part of the public research infrastructure. First, the national scope of the data makes precise estimates of effect sizes possible and meaningful. We learn from this research not only the direction of findings but the magnitude, in national context. For example, Shores and Steinberg (2019) shows that a $1,000 decline in per pupil spending is associated with a 0.17 standard deviation decline in student achievement. Jang and Reardon (2019) predict that a state with between-district income segregation that is 1 standard deviation higher should have a SES-achievement gradient that is 9% larger than the average gradient. And Kuhfeld et al. (2019) show that district achievement correlates .97–.98 across different tests and linking procedures, and growth .82–.90. Data sets like SEDA enable researchers to not only point in a direction, but put a finger on magnitudes, in national context.

Second, all of these articles linked educational data to other variables in other data sets, including the Quarterly Census of Employment and Wages (Shores & Steinberg, 2019), the Transitional Records Access Clearinghouse (Bellows, 2019; Kirksey et al., 2020), the Civil Rights Data Collection (Pearman et al., 2019), the American Community Survey (Jang & Reardon, 2019), and MAP Growth (Kuhfeld et al., 2019). Whether as outcomes, covariates, or question predictors, these merged variables enable researchers to answer important descriptive and causal questions. I hope to see this public research infrastructure continue to expand.

Third, relatedly, all of these articles reflected the value of comparable national educational data over time. SEDA's hidden Rosetta Stone is the essential NAEP, the only assessment that provides multidecade, population-level, and comparable educational outcomes across states. SEDA's first version had only 5 years of data, 2009 to 2013. With the release of Version 3.0 in 2019, there are 8 years of data. Later this year, Version 4.0 will include 10 years of data. As staggered policy implementation is common in the United States, stable national data sets like SEDA should provide ever-expanding opportunities for both descriptive and causal research.

References

Bellows, L. (2019). Immigration enforcement and student achievement in the wake of secure communities. *AERA Open*, *5*(4). https://doi.org/10.1177/2332858419884891

Ho, A. D., & Reardon, S. F. (2012). Estimating achievement gaps from test scores reported in ordinal "proficiency" categories. *Journal of Educational and Behavioral Statistics*, *37*(4), 489–517. https://doi.org/10.3102/1076998611411918

Jackson, C. K., Wigger, C., & Xiong, H. (2018). *Do school spending cuts matter? Evidence from the Great Recession* (No. w24203). National Bureau of Economic Research. https://doi.org/10.3386/w24203

Jang, H., & Reardon, S. F. (2019). States as sites of educational in(equality): State contexts and the socioeconomic achievement gradient. *AERA Open*, *5*(3). https://doi.org/10.1177/2332858419872459

Kirksey, J. J., Sattin-Bajaj, C., Gottfried, M. A., Freeman, J., & Ozuna, C. S. (2020). Deportations near the schoolyard: Examining immigration enforcement and racial/ethnic gaps in educational outcomes. *AERA Open*, *6*(1). https://doi.org/10.1177/2332858419899074

Kuhfeld, M., Domina, T., & Hanselman, P. (2019). Validating the SEDA measures of district educational opportunities via a common assessment. *AERA Open*, *5*(2). https://doi.org/10.1177/2332858419858324

Lafortune, J., Rothstein, J., & Schanzenbach, D. W. (2018). School finance reform and the distribution of student achievement. *American Economic Journal*, *10*(2), 1–26. https://doi.org/10.1257/app.20160567

Pearman, F. A., Curran, F. C., Fisher, B., & Gardella, J. (2019). Are achievement gaps related to discipline gaps? Evidence from national data. *AERA Open*, *5*(4). https://doi.org/10.1177/2332858419875440

Reardon, S. F., Ho, A. D., Shear, B. R., Fahle, E. M., Kalogrides, D., Jang, H., Chavez, B., Buontempo, J., & DiSalvo, R. (2019). *Stanford Education Data Archive* (Version 3.0). http://purl.stanford.edu/db586ns4974

Reardon, S. F., Kalogrides, D., & Ho, A. D. (2019). Validation methods for aggregate-level test scale linking: A case study mapping school district test score distributions to a common scale. *Journal of Educational and Behavioral Statistics*. Advance online publication. https://doi.org/10.3102/1076998619874089

Reardon, S. F., Papay, J. P., Kilbride, T., Strunk, K. O., Cowen, J., An, L., & Donohue, K. (2019). *Can repeated aggregate cross-sectional data be used to measure average student learning rates? A validation study of learning rate measures in the Stanford education data archive* [CEPA Working Paper No. 19-08].

Reardon, S. F., Shear, B. R., Castellano, K. E., & Ho, A. D. (2017). Using heteroskedastic ordered probit models to recover moments of continuous test score distributions from coarsened data. *Journal of Educational and Behavioral Statistics*, *42*(1), 3–45. https://doi.org/10.3102/1076998616666279

Shores, K., & Steinberg, M. P. (2019). Schooling during the great recession: Patterns of school spending and student achievement using population data. *AERA Open*, *5*(3). https://doi.org/10.1177/2332858419877431

Whitehurst, G. J., Chingos, M. M., & Gallaher, M. R. (2013). *Do school districts matter?* Brookings Institution.

Author

ANDREW D. HO is a professor of education at Harvard Graduate School of Education. He is a psychometrician who aims to improve the design, use, and interpretation of test scores in educational policy and practice.