Abstract Title Page

Not included in page count.

Title: Mapping U.S. school district test score distributions onto a common scale, 2008-2013

Authors and Affiliations: Sean Reardon (Stanford University), Demetra Kalogrides (Stanford University), Andrew Ho (Harvard Graduate School of Education),

Abstract Body

Limit 4 pages single-spaced.

Background / Context:

Description of prior research and its intellectual context.

U.S. school districts differ dramatically in their socioeconomic and demographic characteristics (Reardon, Yun, & Eitle, 1999; Stroub & Richards, 2013), and districts have considerable influence over instructional and organizational practices that may affect academic achievement (Whitehurst, Chingos, & Gallaher, 2013). Nonetheless, we have relatively little rigorous large-scale research describing national patterns of variation in achievement across districts, let alone an understanding of the factors that cause this variation. Such analyses generally require district-level test score distributions that are comparable across states. Although these comparisons are possible within states and for selected districts across states,[1] a national district-level database of comparable achievement scores does not exist.

This paper evaluates a linking method applied to a unique national district-level dataset that may achieve the goal of national district-level comparability for research purposes. We begin with a dataset constructed from an application of a method presented at this conference last year (Shear, Castellano, Reardon, & Ho, 2014). The authors fit a heteroskedastic probit model to categorical test score data from the EDFacts Initiative (U.S. Department of Education, 2015), which includes frequencies of students in coarse “achievement levels” from every school district in the U.S. from 2008 to 2012. Data from 2013 are forthcoming. The authors demonstrate that reparameterization of conventional model parameters recover, with remarkable accuracy, means and standard deviations of district test scores from fine-grained data. The method is useful because the latter data are seldom available in practice. However, district means and standard deviations remain incomparable across states, because state test score categories and their underlying test score scales differ.

We employ a linear linking method under the assumption that state test score distributions have the same mean and standard deviation as their counterparts in the National Assessment of Educational Progress (NAEP) for the same subjects, grades, and years. The baseline linking method is reviewed by Kolen and Brennan (2014). Hanushek and Woessman (2012) have employed similar methods for international comparisons. Using NAEP as a basis for linking tests has been deemed infeasible for high-stakes student-level reporting (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999); however, our goal is to support aggregate-level policy analysis. For these purposes, as we demonstrate in this paper, we will treat the issue empirically, using a variety of validation checks.

Purpose / Objective / Research Question / Focus of Study:

Description of the focus of the research.

[1] Limited cross-state district comparisons are possible with the Trial Urban District Assessment (TUDA), which reports scores for 20 large districts bienially on the common scale of the National Assessment of Educational Progress (NAEP). They are also possible with tests that have cross-state district-level adoption, like the Measures of Academic Progress (MAP) test from Northwest Evaluation Association (NWEA); however, participation is nowhere near national. The Common Core State Standards Assessment Consortia may eventually provide more but still not national comparisons.

Our goal is to map district test score means and standard deviations onto a common metric across states and evaluate the accuracy of the mapping.

Setting:
Description of the research location.

The research setting is public elementary and middle school districts in the United States from 2008 to 2013.

Population / Participants / Subjects:
Description of the participants in the study: who, how many, key features, or characteristics.

The population consists of all U.S. public school students in grades 3-8 from 2008 to 2013.

Intervention / Program / Practice:
Description of the intervention, program, or practice, including details of administration and duration.

State testing programs differ but are all mandated under the Elementary and Secondary Education Act to report individual student scores in ordinal proficiency categories in Mathematics and Reading or English Language Arts annually in grades 3-8. States provide these scores by school, grade, and district, to the EDFacts database. States are also mandated to have a sample of their schools participate in the National Assessment of Educational Progress (NAEP) every other year, in Reading and Mathematics, grades 4 and 8. We will have state test score data from 2008 to 2013 and NAEP data from overlapping years, 2009, 2011, and 2013.

Research Design:
Description of the research design.

Please see below.

Data Collection and Analysis:
Description of the methods for collecting and analyzing data.

Analysis of EDFacts data using methods from Shear, Castellano, Reardon, and Ho (2014) yields estimates of district test score means and standard deviations on a common state scale with a state mean of 0 and a state variance of 1. We refer to these estimated district means and standard deviations as $\hat{\mu}^{\text{state}}_{dygb}$ and $\hat{\sigma}^{\text{state}}_{dygb}$, respectively, for district d, year y, grade g, and subject b. These methods also provide estimated standard errors of these estimates, $\hat{\sigma}_{\hat{\mu}^{\text{state}}_{dygb}}$ and $\hat{\sigma}_{\hat{\sigma}^{\text{state}}_{dygb}}$, respectively. We also gather reported reliability statistics from state test score technical manuals, $\rho^{\text{state}}_{sygb}$. Through both the publicly accessible NAEP data explorer and the restricted-use data, we have estimates of NAEP means and standard deviations at the state (s) level, $\hat{\mu}^{\text{naep}}_{sygb}$ and $\hat{\sigma}^{\text{naep}}_{sygb}$, respectively, as well as their standard errors.

Under the assumption that NAEP and state test score means and variances should be the same, we first map district-subgroup means to the NAEP scale linearly, for overlapping years and grades. Because district test score moments are already expressed on a state scale with mean 0 and unit variance, the expressions are simple:

$$\hat{\mu}_{dygb}^{\widehat{\text{naep}}} = \hat{\mu}_{sygb}^{\text{naep}} + \frac{\hat{\mu}_{dygb}^{\text{state}}}{\sqrt{\rho_{sygb}^{state}}} * \hat{\sigma}_{sygb}^{\text{naep}}$$

$$\hat{\sigma}_{dygb}^{\widehat{\text{naep}}} = \hat{\sigma}_{dygb}^{\text{state}} * \hat{\sigma}_{sygb}^{\text{naep}}$$

Note that because district means on the state scale, $\hat{\mu}_{dygb}^{\text{state}}$, are expressed in terms of standard deviation units of the state score distribution, they are attenuated due to measurement error. We disattenuate by dividing the square root of the state test score reliability estimate. Although the district standard deviations, $\hat{\sigma}_{dygb}^{\text{state}}$, are also inflated due to measurement error in an absolute sense, because they are expressed as a percentage of the state standard deviation, which is itself inflated, they need not be adjusted for unreliability. Treating the main terms as independent random variables, we can derive the standard errors of the linked means and standard deviations:

$$\hat{\sigma}_{\hat{\mu}_{dygb}^{\widehat{\text{naep}}}} = \sqrt{\hat{\sigma}^2_{\hat{\mu}_{sygb}^{\text{naep}}} + \frac{1}{\rho_{sygb}^{state}} \hat{\sigma}^2_{\hat{\mu}_{dygb}^{\text{state}}} \hat{\sigma}^2_{\hat{\sigma}_{sygb}^{\text{naep}}} + \frac{1}{\rho_{sygb}^{state}} \hat{\sigma}^2_{\hat{\mu}_{dygb}^{\text{state}}} \left(\hat{\sigma}_{sygb}^{\text{naep}}\right)^2 + \frac{1}{\rho_{sygb}^{state}} \hat{\sigma}^2_{\hat{\sigma}_{sygb}^{\text{naep}}} \left(\hat{\mu}_{dygb}^{\text{state}}\right)^2}$$

$$\hat{\sigma}_{\hat{\sigma}_{dygb}^{\widehat{\text{naep}}}} = \sqrt{\hat{\sigma}^2_{\hat{\sigma}_{dygb}^{\text{state}}} \hat{\sigma}^2_{\hat{\sigma}_{sygb}^{\text{naep}}} + \hat{\sigma}^2_{\hat{\sigma}_{dygb}^{\text{state}}} \left(\hat{\sigma}_{sygb}^{\text{naep}}\right)^2 + \hat{\sigma}^2_{\hat{\sigma}_{sygb}^{\text{naep}}} \left(\hat{\sigma}_{dygb}^{\text{state}}\right)^2}$$

Linkages, on their own, are expressions of wishful thinking, in this case that, had the district been sampled for NAEP, its resulting average score and standard deviation on the NAEP scale would be the linked estimates above. We conduct three additional analyses that assess the validity of the linked estimates for their intended research purposes.

First, NAEP reports scores for 17 state districts (TUDAs) in 2009 and 20 in 2011 and 2013. For these particular large districts, we can compare the NAEP means and standard deviations to their linked means and standard deviations. For each district, we can obtain the discrepancy, $\hat{\mu}_{dygb}^{\widehat{\text{naep}}} - \hat{\mu}_{dygb}^{\text{naep}}$. We report the average of these discrepancies as the bias, and we report the square root of the average squared discrepancies as the Root Mean Squared Error (RMSE). We also report the correlation between the two, as well as a disattenuated correlation that accounts for the standard error of each observation, $\hat{\sigma}_{\hat{\mu}_{dygb}^{\widehat{\text{naep}}}}$.

Second, we have access to a large database of scores from a testing program adopted at the school and district level across the country. In a large number of districts across many states, the number of student test scores exceeds 90% of district's enrollment. We estimate means and standard deviations for these districts, and we report correlations and disattenuated correlations between these and the linked district estimates on the NAEP scale. This correlation is expected to be lower due to any divergence in the construct measured by this third party test and NAEP.

Third, Bandeira de Mello, Bohrnstedt, Blankenship, and Sherman (2015) report mappings of state proficiency standards on the NAEP scale using an equipercentile linkage, based on knowledge of the schools sampled by NAEP and their proficiency rates. We evaluate the alignment between the state EDFacts population and the NAEP population by linearly mapping the estimated cut score on the standardized state test score scale, x_{sygb}^{state}, to the NAEP scale via the same linking function above.

$$x_{sygb}^{\widehat{naep}} = \hat{\mu}_{sygb}^{\text{naep}} + \frac{x_{sygb}^{\text{state}}}{\sqrt{\rho_{sygb}^{state}}} * \hat{\sigma}_{sygb}^{\text{naep}}.$$

We report bias, RMSE, and correlations with the reported mapped standards in the full paper. Discrepancies between these mapped standards and those reported by Bandeira de Mello et al. (2015) are a partial indication of misalignment between the state EDFacts and NAEP populations.[2]

Findings / Results:

Given space limitations, we limit our discussion to findings from the first and most direct recovery benchmark: means and standard deviations of NAEP TUDAs. Table 1 shows bias, RMSE, correlations, and precision-adjusted correlations for the 17-20 TUDAs by subject, grade, and year (EDFacts is due to release 2013 restricted-use data shortly). Figure 1 shows a scatterplot of linked and NAEP-reported means. Precision-adjusted correlations are high. However, persistent positive bias across TUDAs (around a 10th of a NAEP standard deviation) indicates systematic high performance of TUDA districts with their respective state test score distributions, leading to higher-than-expected NAEP mappings.

Possible explanations for these discrepancies include disproportionately higher true performance on state rather than NAEP content relative to other districts, differences in motivation across tests by district, and relative inflation (for example, one district with a positive discrepancy had a known cheating scandal on the state test in one year; the discrepancy reduced by the next administration). Note that the expected bias from a linear mapping is near 0 by design. Thus, the reported RMSE underestimates the variation we would expect if we could perform this analysis on all districts, not just TUDAs. We describe results for other criteria in the full paper.

Conclusions:

A nationwide district-level dataset of means and standard deviations will be a valuable tool for future descriptive and causal analysis if and only if it is valid for its intended research purposes. We use a range of validation approaches, here and in the full paper, to demonstrate that overall recovery as indicated by correlations is strong, but bias, although small, is systematic for certain districts. Following the publication of this paper, we intend to release this dataset to the public, complete with documentation detailing its caveats. We hope to benefit from feedback at the SREE spring conference to ensure that this dataset, unprecedented in its geographical detail, advances valid future research analyses and conclusions.

[2] Incongruence between state and NAEP distributional forms may also explain discrepancies. Incongruence causes linear and equipercentile linkages to diverge but is not a threat to inferences from the linear linkage.

Appendices

Not included in page count.

Appendix A. References

References are to be in APA version 6 format.

Bandeira de Mello, V., Bohrnstedt, G., Blankenship, C., & Sherman, D. (2015). *Mapping state proficiency standards onto NAEP scales: Results from the 2013 NAEP reading and mathematics assessments* (NCES 2015-046). U.S. Department of Education, Washington, DC: National Center for Education Statistics.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M.W., & Hemphill, F. C. (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, DC: National Academy Press.

Hanushek, E. A., & Woessmann, L. (2012). Do better schools lead to more growth? Cognitive skills, economic outcomes, and causation. *Journal of Economic Growth, 17,* 267-321.

Reardon, S. F., Yun, J. T., & Eitle, T. M. (1999). *The changing context of school segregation: Measurement and evidence of multi-racial metropolitan area school segregation, 1989-1995*. Paper presented at the annual meeting of the American Educational Research Association. Montreal, Canada.

Shear, B. R., Castellano, K. E., Reardon, S. F., & Ho, A. D. (2014). *Simultaneous estimation of multiple achievement gaps from ordinal proficiency data*. Paper presented at the spring meeting of the Society for Research on Educational Effectiveness. Washington, DC.

Stroub, K. J., & Richards, M. P. (2013). From resegregation to reintegration: Trends in the racial/ethnic segregation of metropolitan public schools, 1993–2009. *American Educational Research Journal, 50*, 497-531.

U.S. Department of Education. (2015). *EDFacts Submission System User Guide V11.2* (SY 2014-2015). Washington, DC: EDFacts. Retrieved from http://www.ed.gov/edfacts

Whitehurst, G. J., Chingos M. M., & Gallaher, M. R. (2013). *Do School Districts Matter?* Washington, DC: Brookings Institution.

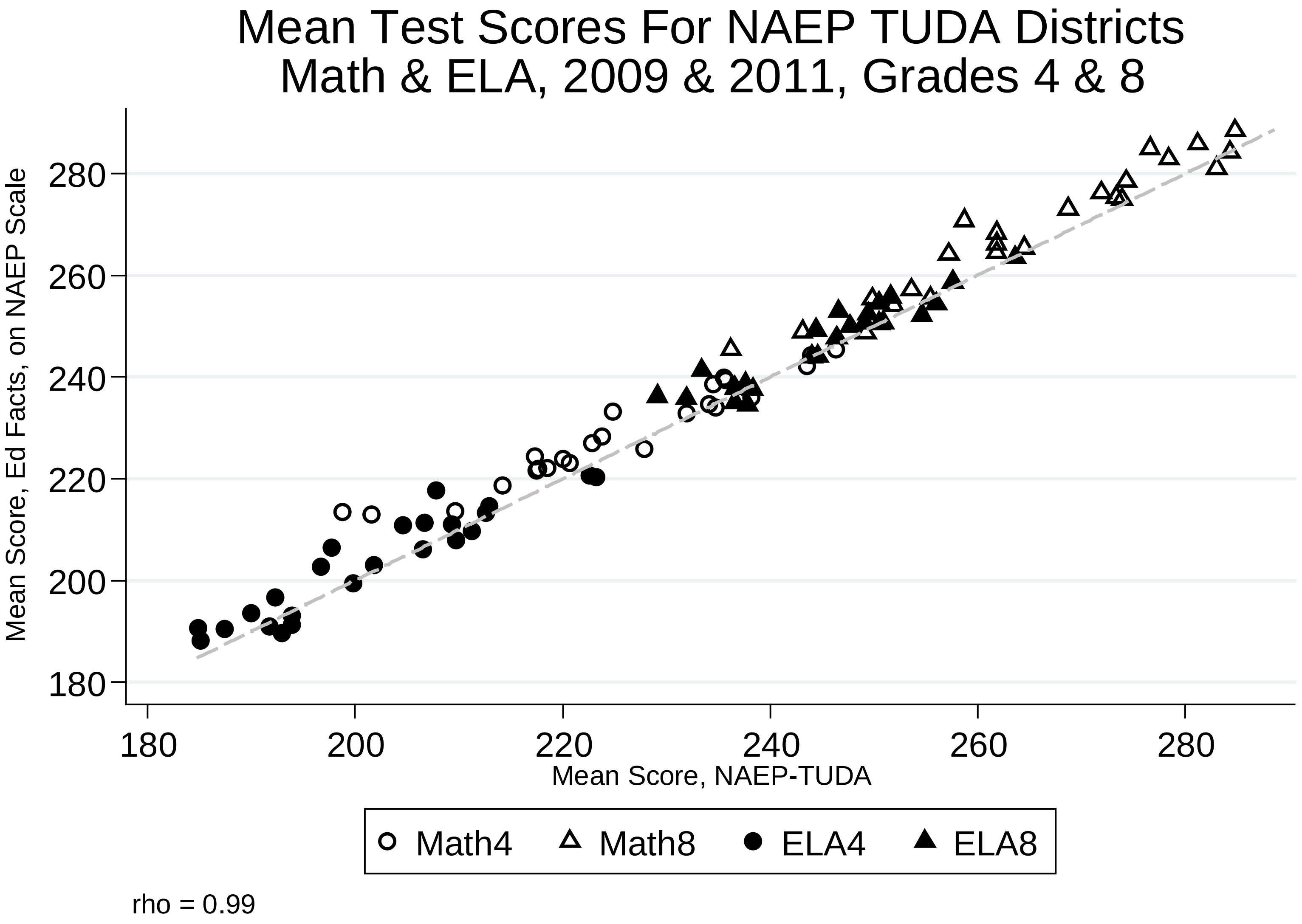
Appendix B. Tables and Figures

Not included in page count.

Table 1: Estimated correlations between NAEP-linked EdFacts estimates and NAEP TUDA estimates

			Linear Calibration			Precision-Adjustment			Linear Calibration-EPE			Precision-Adjustment- EPE		
Subject	Grade	Year	Correlation	Bias	RMSE	SE_State	SE_NAEP	Adj. Corr.	Correlation	Bias	RMSE	SE_State	SE_NAEP	Adj. Corr.
Reading	4	2009	0.93	1.40	4.10	1.57	1.65	0.94	0.94	2.18	4.42	1.88	1.67	0.96
		2011	0.94	0.80	4.48	1.31	1.54	0.96	0.94	1.61	3.78	1.37	1.59	0.97
	8	2009	0.87	1.63	4.19	1.42	1.67	0.89	0.93	1.86	3.73	1.64	2.13	0.97
		2011	0.95	1.26	3.25	1.13	1.30	0.97	0.95	1.46	2.98	1.17	1.27	0.97
Math	4	2009	0.95	3.77	5.16	2.46	1.23	0.95	0.96	4.06	5.90	3.40	1.26	0.97
		2011	0.95	2.56	4.87	1.05	1.02	0.95	0.97	3.04	4.74	1.10	1.01	0.98
	8	2009	0.92	4.28	5.98	1.33	1.37	0.92	0.96	4.74	6.15	1.40	1.44	0.97
		2011	0.93	3.01	5.18	1.13	1.25	0.94	0.98	3.25	3.96	1.19	1.27	0.99

Figure 1: NAEP-linked EDFacts and NAEP TUDA estimated means, grades 4-8, 2009 and 2011.


Mean Test Scores For NAEP TUDA Districts
Math & ELA, 2009 & 2011, Grades 4 & 8
Mean Score, Ed Facts, on NAEP Scale
Mean Score, NAEP-TUDA
○ Math4
△ Math8
● ELA4
▲ ELA8
rho = 0.99