Pooling Standardized Test Scores Across States: Using NAEP to Standardize and Rescale

Author: Ying Liu (presenting author), Janet Li, Anna Saavedra, Shira Korn

Background: Studies evaluating program implementation across multiple states are common. State standardized test scores often serve as low-cost measures of student baseline and outcome achievement. Since each U.S. state designs standardized assessments independently to meet unique learning standards, scores across states are typically incomparable (Chingos & Blagg, 2018; Olsen et al., 2012). Researchers need a way to pool student-level scores across states that minimizes error or bias induced by such inequivalence. May et al (2009) suggest estimation of impacts on state-specific subsamples and combining results using meta-analysis techniques. However, this strategy requires large samples within each state, and researchers trained in meta-analysis—in concert, an uncommon condition. "Rescaling" student-level data to a common metric and analyzing the data as if it came from the same source is a more powerful and efficient alternative.

Objective: Our objective was to rescale eighth-grade state standardized English Language Arts (ELA) and mathematics test scores from across five states, adjusting for differences between states and across years to minimize bias or error induced by inequivalent test scores.[1] Our "two-step" approach was to 1) standardize state test scores within states using respective statewide means and standard deviations (SD), and then 2) adjust for between-state differences using the states' National Assessment of Educational Progress (NAEP) scores.

Two assumptions justified our approach. First, state and NAEP assessments similarly reflect student's overall performance at the population level. Second, studies linking state and NAEP assessments at the student level suggest a correlation around 0.75 across assessment types (Thissen, 2007), comparable to typical test-retest reliability. Reardon, Kalogrides & Ho (2017) similarly used NAEP scores to realign state-specific assessment distributions onto a common metric.

Hypothesis: We hypothesized that: 1) The eighth-grade state scores rescaled using our two-step approach would strongly predict students' performance on college entrance exams (i.e. PSAT, SAT, and/or ACT scores), and 2) Our approach would outperform prediction of college entrance examination performance relative to eighth-grade state scores rescaled through two alternative approaches: a) standardizing against statewide norms without adjusting for between-state difference ("state-norm only" approach), and b) standardizing against sample distributions ("sample statistics" approach).

Participants: To inform an evaluation of a curriculum intervention on students' academic performance, we collected student-level eighth-grade state standardized test scores from districts located in five states, as well as college entrance exam scores for the same students. The 2,560 sample students were 39 percent White, 34 percent Hispanic, 16 percent Asian, and 8 percent Black, with 40

[1] Our objective was *not* to produce scores comparable and accurate enough to rank order students, for example, to compare performance of a student in one state to the performance of a student in another state.

percent eligible for free-or-reduced price lunch. All had at least one mathematics or ELA eighth-grade state test score; 1,920 (75 percent), also had at least one national assessment score (PSAT, SAT or ACT).

Research Design:

"Two-step" calculation

Our first step was to standardize each student's scale score by the statewide norm of the test administration, for the specific subject and in the given year. For example, we standardized students' 2015 eighth-grade California Assessment of Student Performance and Progress (CAASPP) mathematics tests using the 2015 CAASPP mathematics state mean score and SD. Each student's "z-score" represented his/her relative standing in the population of each test administration. The second step was to rescale each student's z-score using their state's NAEP mean and SD in the given test year. Continuing the CAASPP mathematics example, we obtained California's state mean score and SD on the 2015 administration of the NAEP mathematics assessment, then used these norms to scale up students' z-scores obtained in step 1. Let $X_{j,s}$ be the original scale score for person j and state s, μ_s and σ_s be the population mean and SD for the state assessment scores, and μ_s^{NAEP} and σ_s^{NAEP} for NAEP scores. We computed rescaled score $X_{j,s}^*$ as follows.

$$\text{Step 1: } Z_{j,s} = \frac{X_{j,s} - \mu_s}{\sigma_s}$$

$$\text{Step 2: } X_{j,s}^* = Z_{j,s} * \sigma_s^{NAEP} + \mu_s^{NAEP}$$

As alternatives to our two-step approach, we also standardized the students' state test scores against a) statewide norms without adjusting for between-state difference ("state norm only"), and b) sample distributions ("sample statistics").

Predictive utility of "two-step" relative to "state-norm only" and "sample statistics" approaches

We used students' two-step rescaled eighth-grade scores to predict their subsequent performance on national assessments. To examine the extent to which the rescaled scores aligned on a common metric, we conducted an F-test comparing (a) a full model in which intercepts and slopes varied across source of state tests, to (b) a null model forcing intercepts and slopes to be the same. We interpreted the magnitude of the F-statistic as the "loss" in predictability due to forced equation of intercepts or slopes. Larger F-statistic values imply larger across-state difference left unaccounted for *after rescaling*. Inequivalent intercepts signal unresolved between-state difference in achievement levels, whereas inequivalent slopes suggest unequal unit size across state test scores.

We then used the same analytic approach to assess the predictive utility of the "state-norm only" and "sample statistics" rescaling approaches.

Results: Our two-step approach resulted in correlations between eighth-grade test scores and the national test scores of 0.69 for mathematics and 0.71 for ELA (Table 1). F-tests revealed nonsignificant or minor unexplained differences across state tests. These results imply that our two-step rescaling method is effective.

Table 1. Comparison across different rescaling methods.

Subject	Rescaled through…	Correlation	F-statistic for Common Intercept	F-statistic for Common Slope
Math	Two-step	0.69	1.4	2.1
	State Norm Only	0.68	25.0***	2.0
	Sample Statistics Only	0.62	126.1***	4.1**
ELA	Two-step	0.71	2.5*	2.8*
	State Norm Only	0.72	42.0***	3.3**
	Sample Statistics Only	0.60	206.1***	15.1***

Notes: * p<0.05, ** p<0.01, *** p<0.0001; F-statistic degrees of freedom (5, 1896).

Rescaling using the alternative "sample statistics" approach resulted in lower correlations. The correlations obtained through the "state-norm only" approach were similar to those realized through our two-step approach. However, the F-statistics were smallest for our approach, larger for the "state norm only" approach, and largest for the "sample statistics" approach. The difference was particularly evident when equating intercepts, suggesting that our two-step approach was considerably better at accounting for cross-state differences in achievement levels.

Conclusion: Cross-state studies often require pooling of standardized test scores across states. Standardizing against state mean scores and SD's and then adjusting for between-state differences using NAEP state-level sample statistics is more effective than standardizing against state norms or sample statistics. State and NAEP summary statistics are publicly available, making this method broadly accessible.

References:

Chingos, M. M. & Blagg, K. (2018). A better way to compare state performance on NAEP [Blog Post]. *Education Next*. Retrieved from https://www.educationnext.org/better-way-compare-state-performance-naep/

May, H., Perez-Johnson, I., Haimson, J., Sattar, S., & Gleason, P. (2009). *Using state tests in education experiments: a discussion of the issues* (NCEE 2009-013). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Olsen, R. B., Unlu, F., Jaciw, A. P., and Price, C. (2012). *Estimating the impacts of educational interventions using states tests or study-administered tests* (NCEE 2012- 4016). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Reardon, S.F., Kalogrides, D., & Ho, A. (2017). *Linking U.S. school district test score distributions to a common scale* (CEPA Working Paper No.16-09). Retrieved from Stanford Center for Education Policy Analysis: http://cepa.stanford.edu/wp16-09

Thissen, D. (2007). Linking assessments based on aggregate reporting: background and issues. In N.J. Dorans, M. Pommerich, & P.W. Holland (Eds.) *Linking and aligning scores and scales*(Pp. 287-312). New York, NY: Springer.