USC Dornsife
Center for Economic and Social Research

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

Ying Liu, Janet Li, Anna Saavedra, Shira Korn

Acknowledgements
with gratitude, we acknowledge financial support for this research from the International Baccalaureate Organization. Any flaws remaining are solely the authors' responsibility. Findings and conclusions expressed in this presentation are those of the authors and do not necessarily reflect the views of the International Baccalaureate Organization.
Contact: Ying Liu, liu.ying@usc.edu

BACKGROUND

Studies evaluating program implementation across multiple states are common. State-wide standardized test scores are often low-cost measures of student baseline and outcome in K-12 context. Since each U.S. state designs standardized assessments independently to meet unique learning standards, scores across states are typically incomparable.

May et al (2009) suggest to estimate impact on state-specific subsamples and then combine results using meta-analysis techniques. However, this requires large samples per state and researchers trained in meta analysis.

Is there a way to "rescale" student-level test score to a common metric so the data can be analyzed as if it came from the same source?

OBJECTIVE AND APPROACH

Rescale eighth-grade state standardized English Language Arts (ELA) and mathematics test scores from students across five states, adjusting for differences between states and across cohort years to minimize bias or error induced by inequivalent test scores.

"Two-step" approach

(1)Standardize scores within states per test year using respective statewide mean and standard deviations (SD).

(2)Adjust for between state, across year differences using the states' National Assessment Education Progress (NAEP) scores.

Assumption/Justification: (a) state and NAEP assessments similarly reflect students' performance at the population level; (b) studies linking state and NAEP assessment scores suggest a correlation around 0.75 (Thissen, 2007). Reardon et al. (2017) similarly used NAEP scores to realign state-specific assessment score distributions onto a common metric.

Hypothesis

(1)The eighth-grade state scores rescaled using our two-step approach would strongly predict students' performance on college entrance exams.

(2)Our approach would outperform two alternative approaches in such prediction: (a) standardizing against statewide norms without adjusting for between-state difference, and (b) against sample distribution.

PARTICIPANTS

To inform an evaluation of a curriculum intervention, we collected student-level scores on eight-grade state exams and PSAT/SAT/ACT across five states. N=2560, 39% White, 34% Hispanic, 16% Asian, & 8% Black, with 40% eligible for free-and-reduced price lunch.

RESEARCH DESIGN

"Two-step" calculation

(1) Standardize each student's scale score by the statewide norm of the test administration, for the specific subject and in the given year. The z-score represented the student's relative standing in the population of that test administration.
(2) Rescale each student's z-score with their state's NAEP mean and SD in the given test year.

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.

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

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

Alternative approaches

(1) Standardize against statewide norm without adjusting for between-state difference ("state norm only" approach).
(2) Standardize against sample distribution ("sample statistic" approach).

PREDICTIVE UTILITY OF RESCALED SCORES

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.

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.

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).

CONCLUSIONS AND DISCUSSION

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.

Future direction

(1) Evaluate to what extent scores rescaled through different methods result in different or biased impact estimate.
(2) Here we only rescaled the cross-sectional test data. To what extent the method is applicable to longitudinal data to assess growth requires further investigation.

Q & A:
Is this replacement of linking studies?

- No!
- 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.

Q & A:
Why use high-stake college entrance exams for predictive utility?

- Literature shows decent correlation between scores on state accountability exams and scores on college entrance exams. For example
  - Correlation ranges from 0.55-0.74 between PARCC and SAT/ACT (Maryland Assessment Research Center, 2016)
  - Correlation ranges from 0.61-0.82 between SBAC and SAT (Kurlaender, et al., 2018)
- Lack of alternatives
  - State exam scores on a later grade level seem to be a natural criterion. However, it suffers from the same issue of not being comparable across states.
  - GPA? Even less standardized or comparable than the state exam scores…
  - Other standardized K-12 assessments such as NWEA's MAP? Less prevalent than PSAT/SAT/ACT and many students do not take.
  - Future performance post-secondary? Need to track students; unattenable.
- Acknowledge the potential misalignment on test content, purpose of use, etc.
  - The college entrance exam may not be considered substitute for the state exams to evaluate student's individual performance or growth.

Q & A:
Why not simply account for the test difference using state fixed effect?

- The test difference is more than just the difference across states (e.g., MA vs. CA).
  - Difference across years of administration
  - Many states transit to new assessment systems (e.g., MA and IL transits from MCAR and ISAT to PARCC) and usually no concordance available across old and new systems.
  - Some states use end-of-course (EOC) assessment framework instead of end-of-grade (EOG). Assessment on, say, mathematics, could consist of a variety of subjects such as general math, algebra I & II, geometry, for the same cohort of students assessed in the same year.

Reference

- Chingos, M. M. & Blagg, K. (2018). A better way to compare state performance on NAEP [Blog Post]. *Education Next*.
- Kurlaender, M., Kramer, K. A., & Jackson, E. (2018). *Predicting college success: How do different high school assessments measure up?* Research Paper Series by Policy Analysis for California Education (PACE).
- Maryland Assessment Research Center. (2016). *The relationship between the PARCC test scores and the college admission tests: SAT/ACT/PSAT.* Maryland State Department of Education.
- 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 ofeducational 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).
- 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.