IES Institute of Education Sciences

APPENDIXES
February 2023

Regional Educational Laboratory Mid-Atlantic
At Mathematica

Stabilizing Subgroup Proficiency Results to Improve the Identification of Low-Performing Schools

See https://ies.ed.gov/ncee/rel/Products/Publication/106926 for the full report.

Appendix A. Literature review

The Regional Education Laboratory (REL) Mid-Atlantic study team used Bayesian hierarchical modeling to stabilize school accountability data, reducing measurement error that results from small sample sizes. Bayesian modeling achieved this goal by considering each school's or subgroup's data in the broader context of data from other schools or subgroups, as well as historical data from the same school or subgroup. Structured assumptions about these relationships–for example, the assumption that schools' test scores are likely to change at similar rates over time–strengthen estimates by reining in extreme values that most often arise by chance. It is well established in statistics that stabilization based on such assumptions–also known as partial pooling, shrinkage, or reliability adjustment–predicts future performance better than unstabilized alternatives (Efron & Morris, 1977). Specifically, an estimate of school performance that is stabilized by drawing on information about other schools more accurately predicts that school's future performance than does an estimate based only on that school's recent performance.

Stabilization has long been applied to performance measurement, both in health care (in hospital quality-measurement programs) and in education (in teacher evaluation). In a Centers for Medicare & Medicaid Services report on hospital performance measurement, the Committee of Presidents of Statistical Societies endorsed the use of Bayesian stabilization for hospital-specific measures, such as the readmission or mortality rate (Ash et al., 2011). The authors argued that stabilization is necessary because it both reduces variation across entities that reflects random error rather than true differences in performance and diminishes the impact of regression to the mean on an entity's performance trajectory over time.

Ash et al. (2011) acknowledged that stabilization has a greater effect on hospitals with few patients, an idea that Herrmann et al. (2016) explore in the context of value-added modeling for teacher evaluation: they hypothesized that teachers with few students, or with students whose achievement is hard to predict, may be assigned consequences at different rates than other teachers, primarily because their performance will be affected more by the stabilization process at the core of value-added models. However, in an empirical analysis of student and teacher data from the District of Columbia Public Schools in the 2011/12 school year, Herrmann et al. (2016) found that stabilization had little effect on whether such teachers were assigned consequences.

Taken together, these studies indicate that stabilization has great potential to improve the reliability of accountability data—by reducing measurement error and thereby enhancing predictive ability—without unduly influencing the outcomes of accountability calculations. It is possible to use more traditional, frequentist approaches to perform stabilization. However, Bayesian inference more completely accounts for the sources of uncertainty that contribute to the model, such as the variation across changes in a school's proficiency rates over time. Frequentist inference treats some of these sources of variation as fixed, resulting in overly confident estimates (Vasishth et al., 2018). Fully Bayesian inference also permits the framing of results in intuitive, probabilistic terms, such as "there is a 78 percent probability that the proficiency rate for economically disadvantaged students in school X is below 30 percent." Frequentist inference does not permit such statements.

In this study, the study team directly examined whether stabilization increased the reliability of academic proficiency rates used in school accountability calculations, as well as how much the set of schools identified for additional support was affected by replacing unstabilized proficiency rates with stabilized rates in the accountability system.

References

Ash, A.S., Fienberg, E., Louis, A., Norm, S.L.T., Stukel, A., & Utts, P.J. (2011). *Statistical issues in assessing hospital performance*. Commissioned by the Committee of Presidents of Statistical Societies for the Centers for Medicare & Medicaid Services. Centers for Medicare & Medicaid Services.

Efron, B., & Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, *236*(5), 119–127.

Herrmann, M., Walsh, E., & Isenberg, E. (2016). Shrinkage of value-added estimates and characteristics of students with hard-to-predict achievement levels. *Statistics and Public Policy*, *3*(1), 1–10. https://doi.org/10.1080/2330443X.2016.1182878.

Vasishth, S., Mertzen, D., Jager, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language, 103*(1), 151–175.

Appendix B. Data and methods

The Pennsylvania Department of Education (PDE) provided school-level data for this study. The dataset contains one record for each combination of school and subgroup for each school year from 2015/2016 through 2018/2019, for all elementary, middle, and high schools included in PDE's accountability system in those years. In Pennsylvania, students in grades 3 through 8 are tested each year, and high school students are tested at the end of specific courses. For each school, the analytic dataset included all tested students across courses and grade levels in the following subgroups: racial/ethnic categories (Asian students, Black students, Hispanic students, White students, multiracial students);[1] economically disadvantaged students; students with disabilities; and English learner students. Key variables included the percentage of students assessed as proficient or advanced on math and English language arts (ELA) assessments in each school-subgroup combination and the number of tested students in each school-subgroup combination in each year.

Additionally, PDE provided publicly available files used for annual meaningful differentiation. These files contain the values of the remaining five accountability indicators[2] for each school, subgroup, and year, as well as whether each school and subgroup was identified for Targeted Support and Improvement (TSI) or Additional Targeted Support and Improvement (ATSI).

The study team fit two Bayesian models: one intended to mirror PDE's ATSI accountability process, the other intended to mirror PDE's TSI accountability process. The team fit each model separately to data from each subgroup to avoid artificially inflating the precision of the estimates by treating subgroups as independent when they are likely to contain overlapping students.

- **ATSI model:** A cross-sectional model that borrows strength across schools, using the two-year (2016/17 and 2017/18) average proficiency rates that PDE applied in its most recent ATSI accountability calculations. In this model, the dataset contains one observation per school-subgroup combination. The model is:

$$\bar{y}_j \sim N\left(\alpha_0 + \alpha_j, \frac{\sigma^2}{\bar{n}_j}\right).$$

  The model uses the following data:

  - $\bar{y}_j$ represents the two-year average proficiency rate for the given subgroup in school j over the data period that PDE uses for ATSI determinations (2016/17 and 2017/18), centered and standardized to have a mean of 0 and a standard deviation of 1.
  - $\bar{n}_j$ is the average number of tested students in the given subgroup over the two-year data period that PDE uses for ATSI determinations (2016/17 and 2017/18). More specifically, $\bar{n}_j$ is the sum of the number of test-takers in math and ELA over the two years, divided by two, because in a single year most of the students who took the math assessment also took the ELA assessment.

- **TSI model:** A longitudinal model that borrows strength both across schools and over time within a school to estimate proficiency rates for each school, using four years of average data (2015/16 through 2018/19). The model is:

$$y_{jt} \sim N\left(\alpha_0 + \alpha_j + \beta_0 t + \beta_j t + \gamma P_{jt}, \frac{\sigma^2}{n_{jt}}\right).$$

[1] Native American/Alaska Native students and Hawaiian/Pacific Islander students were excluded from the analysis because of insufficient data.

[2] The other five indicators are academic growth, progress toward fluency for English learner students, career readiness, regular attendance, and graduation rates.

The model uses the following data:

- y_{jt} represents the proficiency rate for the given subgroup in school j in year t, centered and standardized to have mean of 0 and standard deviation of 1.
- t represents numeric indicators for each year, where $t = 0$ corresponds to the 2015/16 year, $t = 1$ the 2016/17 year, and so forth. The year indicators are then centered so that $t = 0$ corresponds to the middle year, to improve the interpretability of the intercept term.
- P_{jt} is an indicator that school j was penalized in year t for low participation in state assessments.[3]
- n_{jt} is the number of tested students, or the number of students in the denominator if the 95 percent participation penalty policy applies,[4] in the given subgroup in school j in year t. The minimum value for n_{jt} is 10; school-subgroup-year combinations with smaller n_{jt} are excluded from the analysis, as well as from PDE's TSI and ATSI determinations.

Proficiencies are calculated as the weighted average of the math and ELA proficiencies, and numbers of tested students are averaged across math and ELA.

Model parameters are defined as follows:

- σ^2 represents residual error.
- α_0 is an overall intercept term for the scaled and centered proficiency rates.
- α_j is a school-specific intercept representing the difference between the overall proficiency level at school j and the overall proficiency level across schools, α_0. This term also accounts for nonindependence across observations of the same school in different years.
- β_0 is an overall slope describing the rate of change in proficiency over time across all schools.
- β_j is a school-specific slope representing the difference between the rate of change in proficiency at school j and the overall rate of change in proficiency across schools, β_0.
- γ is the average difference in proficiency rates between schools that are penalized due to low participation in state assessments and those that are not penalized.

The ATSI model stabilizes the two-year average proficiency rates using information across schools to reduce measurement error in each school's estimated two-year proficiency rate. Because only one two-year proficiency rate was available (PDE assigned schools to ATSI in 2018 using data from the 2016/17 and 2017/18 academic years), this model does not include historical data. The inclusion of time trends using historical two- or three-year proficiencies will be possible as more data become available; however, the cross-sectional approach may continue to be preferable if the decision rules that determine how PDE calculates accountability indicators change, disrupting the time series.

The TSI model takes advantage of historical information about each school's single-year proficiencies to improve its estimates of the schools' trajectories. Although the model makes a simplifying linear assumption, this simple linear form is more appropriate for the short time series–just four time points, 2015/16, 2016/17, 2017/18, and

[3] The study team also considered a continuous formulation where γ represents the correlation between the percentage of students participating in testing at a school and the school's proficiency rate. However, exploratory data analysis did not support a linear relationship, so the team proceeded with the binary formulation.

[4] The denominator is equal to either the sum of the number of students tested in math and ELA or 95 percent of the sum of the numbers of students eligible for testing in those subjects, whichever is higher.

2018/19—than a more complex model would be. More complex models would be possible with a longer time series, but the simpler linear specification may be preferable if PDE anticipates changes to the decision rules for calculating accountability indicators that would disrupt the time series.

Both models use standardized proficiency rates (transformed to *z*-scores by subtracting the mean and dividing by the standard deviation) so that the response variable has a mean of 0 and unit variance. Standardizing the outcome helps to satisfy the assumptions of a normal likelihood, such as the presence of both negative and positive outcome values, and makes it possible to use recommended default prior distributions. After fitting the models, the study team back-transformed the stabilized proficiency outcomes to the 0–100 scale by multiplying by the standard deviation and adding the mean. Because the Bayesian model stabilizes by shrinking estimates toward the overall mean (both ASTI and TSI models) and the overall time trend (TSI model only), the back-transformed stabilized proficiency rates are within the 0–100 range, even though the normality assumption would allow them to take on values outside this range.

To fit a regression model in the Bayesian paradigm, it is necessary to specify not only the regression equation but also prior distributions that represent the range of likely values for each parameter to be estimated. These distributions encode assumptions about the magnitudes of parameters and about how parameters relate to one another. For this study, the team used the following prior distributions:

$$\alpha_0, \beta_0, \gamma \sim N(0,1)$$

$$\alpha_j \sim N(0, \sigma_a^2)$$

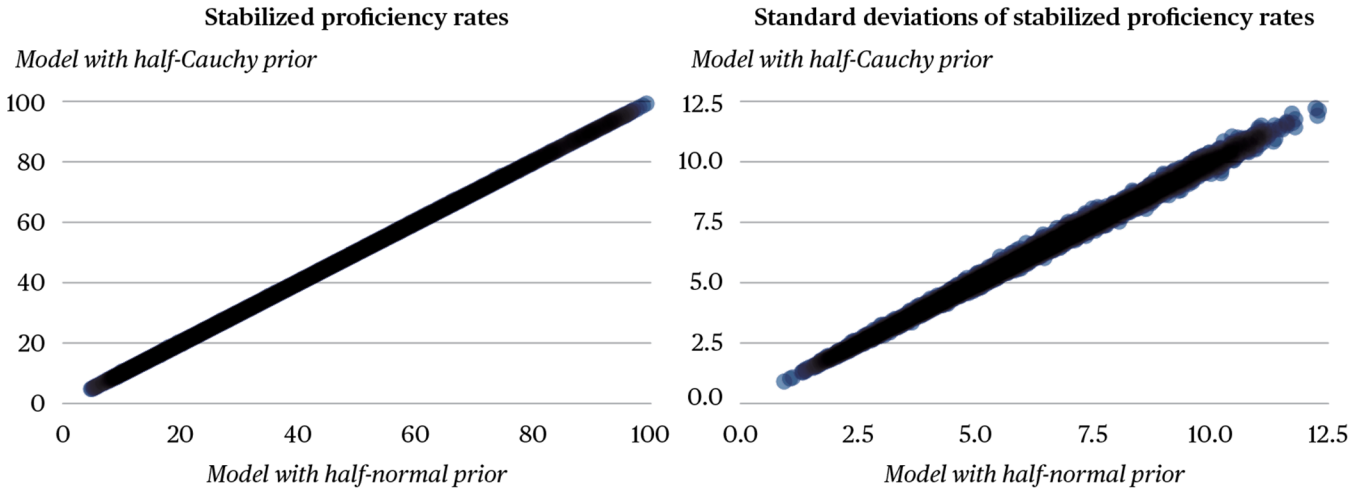
$$\beta_j \sim N(0, \sigma_b^2)$$

$$\sigma, \sigma_a, \sigma_b \sim N^+(0,1).$$

The prior distributions for $\alpha_0, \beta_0, \gamma,\ \sigma, \sigma_a,$ and σ_b represent gentle assumptions that these parameters are unlikely to be large; the superscript plus in the prior distributions for $\sigma, \sigma_a,$ and σ_b indicates that these parameters can take on only positive values. The prior distributions for α_j and β_j induce stabilization across school-specific intercepts and slopes by assuming, for example, that the average proficiency levels in each school come from a common normal distribution. The magnitude of the variance of this distribution, σ_a^2, determines how much the model stabilizes across schools. The study team also estimated σ_a^2 from the data, so that the amount of variation across schools in average proficiency dictates the amount of stabilization the model performs.

For the ATSI model, the study team conducted sensitivity analysis using Cauchy$^+$(0,1) prior distributions for σ and σ_a and did not find that the stabilized proficiencies are sensitive to the choice of prior distributions (figure B1).

Figure B1. Stabilized proficiencies were not sensitive to the choice of Bayesian priors for model parameters σ and σ_a



Note: Data are the stabilized two-year average academic proficiency rates (left) and their standard deviations (right) for a given school and subgroup for the combined 2016/17 and 2017/18 academic years.
Source: Pennsylvania Department of Education data.

As noted, the study team fit each model separately to data from each subgroup to avoid artificially inflating the precision of the estimates by treating subgroups as independent when they are likely to contain overlapping students. The team fit models using Hamiltonian Monte Carlo as implemented in the Stan probabilistic programming language (Stan Development Team, 2021), via its R interface, rstan. The team assessed convergence and mixing using the Gelman-Rubin diagnostic and effective sample sizes; all Gelman-Rubin statistics for parameters used to compute fitted values were within the accepted range (0.9-1.1), and all but a very few were within the preferred range of 0.95-1.05. Similarly, across fits, all model parameters used to calculate fitted values had effective sample sizes of roughly 50 or more. In each model fit, three-quarters or more of the parameters had 500 or more effective samples, and half of the parameters had more than 3,000 effective samples. Although some of these diagnostics fall shy of the standards recommended in the literature, the study team chose to proceed because, after refitting the models several times to improve diagnostics, there was very little sensitivity in the results to the number of Monte Carlo iterations run.

After fitting both models to data from each subgroup, the team used the fitted values from each model–also called the stabilized proficiency rates–to answer the research questions. The fitted values were calculated as the mean of the posterior distribution of the linear predictor for each observation in the dataset. Although the study team calculated the posterior standard deviations and the 95 percent credible interval bounds for the fitted values, this information was not incorporated into further analysis, to parallel the information available for unstabilized proficiency rates and to mirror PDE's current practice of not accounting for statistical uncertainty in accountability calculations. Future research could explore ways of incorporating this information into the accountability system, for example, by calculating the probability that each school meets the performance cutoff for each accountability measure or by calculating the probability that each school meets the standards for ATSI or TSI identification.

Research question 1

For the ATSI model, the study team compared the stabilized and unstabilized two-year average proficiency rates for the 2016/17 and 2017/18 academic years. For the TSI model, the study team compared the stabilized and unstabilized proficiency rates for the 2018/19 year. Although the TSI model was fit using four years of data–

starting with the 2015/16 academic year—the study team focused on the effects of stabilization in the 2018/19 academic year to align with PDE's timeline for TSI identification. TSI schools were identified for the first time in 2019, using the previous academic year's data.

Research question 2

For both the ATSI and TSI models, the study team examined how stabilization moderates the relationship between the variability in proficiency rate estimates and sample size. The study team assessed this relationship visually, using scatter plots, and quantitatively. In the quantitative analysis, the study team divided the data into categories based on the number of tested students in the school-subgroup: 10–19, 20–29, 30–49, 50–99, 100–199, and 200–500. For the unstabilized and stabilized proficiencies, the team first took the standard deviation of the proficiency for each combination of subgroup and sample size category. The study team then took the median and interquartile range of the subgroup-specific standard deviations for each sample size category. The study team compared the relationship between standard deviation and sample size for the stabilized and unstabilized estimates separately for the results of the ATSI and TSI models.

Research question 3

The study team focused on ATSI for this research question because the academic proficiency cutoff for TSI identification varies depending on the subgroup's academic growth; in addition, a TSI identification has only a minimal impact on schools' operations.

In practice, when determining which schools qualify for ATSI, PDE uses discretion to adjudicate difficult cases. The study team would not be able to replicate these judgments, so to facilitate a head-to-head comparison between the ATSI determinations based on stabilized and unstabilized proficiency rates, the team implemented PDE's accountability rules as an algorithm. The study team then applied this algorithm twice, once using unstabilized proficiency rates and once using stabilized proficiency rates. The ATSI identifications based on unstabilized proficiency rates were considered as the baseline for the comparison. The study team then determined how many schools changed ATSI status—from identified to not identified, and vice versa—when the stabilized proficiency rates were used.

Reference

Stan Development Team. (2021). *Stan modeling language users guide and reference manual, 2.30*. Retrieved December 19, 2021, from https://mc-stan.org.

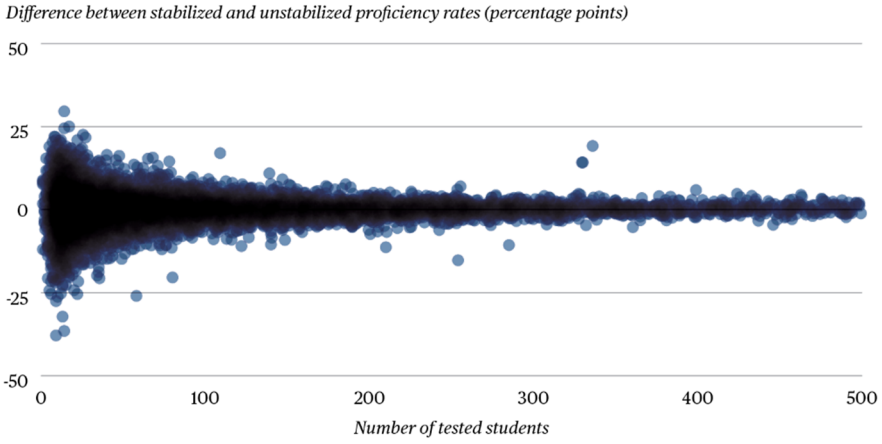
Appendix C. Supplemental results

This appendix presents the results for the Targeted Support and Improvement (TSI) model.

Stabilization had the greatest effect on small school-subgroup combinations

Stabilized proficiency rates differed from unstabilized rates particularly for small school-subgroup combinations. Figure C1 shows the difference between unstabilized and stabilized annual proficiency rates by sample size. The figure has a characteristic funnel shape, with a wider range of differences among smaller school-subgroup combinations (left side of figure) than among larger combinations (right side). This pattern aligns with the expectation that stabilization is most influential for small school-subgroup combinations, where small sample size increases measurement error and correspondingly decreases reliability. Larger school-subgroup combinations were minimally affected. In all cases, the stabilized proficiency rates were closer to the mean proficiency rate in the subgroup than the unstabilized rates were, indicating that stabilization shifted the estimates in the expected direction.

Figure C1. Stabilization was more influential for smaller subgroups (fewer than 100 tested students) than larger ones for annual proficiency rates in the Targeted Support and Improvement model



Note: Each data point represents average academic proficiency rates for a given combination of school and subgroup in the 2018/19 academic year. The horizontal axis represents the number of tested students in that school, subgroup, and year, and the vertical axis represents the difference between the stabilized and unstabilized proficiency rates for that school, subgroup, and year. The funnel shape of the figure, with greater dispersion on the left than on the right, indicates that stabilization affects smaller schools more than larger ones, in line with theory.
Source: Pennsylvania Department of Education data.

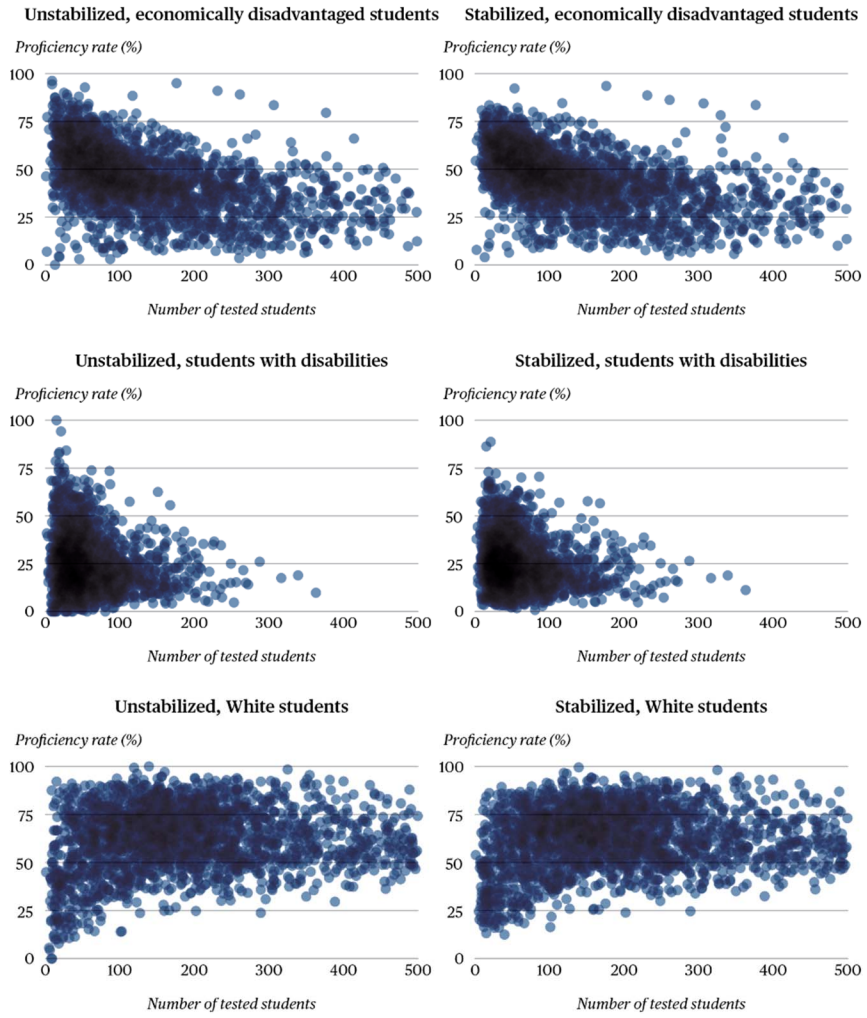
For TSI, the average amount of stabilization is 3.3 percentage points in both directions, with stabilization of less than 4.5 percentage points for 75 percent of the subgroups.

Comparing figure C1 for the TSI model with figure 1 in the main report for the Additional Targeted Support and Improvement (ATSI) model suggests that the ATSI model had more influence on the stabilized estimates. Because the input data for the ATSI model, as two-year average proficiency rates, were somewhat more stable than the annual proficiency rates used in the TSI model, less stabilization might be expected in the ATSI model rather than more. However, a model's degree of stabilization is a function of the relative precision of the data; a dataset with a wider range of precisions, in this case sample sizes, will stabilize more than a dataset with a narrower range of precisions. The two-year average proficiency rates used to fit the ATSI model had a markedly wider range of precisions because the corresponding sample sizes reflect the number of tested students over two years of state assessments, as compared with one year of assessments in the TSI model. In this case, then, the slightly greater degree of stabilization in the ATSI model aligned with expectations.

Stabilization moderated the relationship between subgroup size and variation in proficiency rates, indicating an improvement in statistical reliability

Similar to the ATSI results, the unstabilized TSI data also showed a characteristic funnel pattern: proficiency rates varied more among smaller schools and subgroups than among larger schools and subgroups. Figure C2 depicts the relationship between sample size and variability for both unstabilized and stabilized proficiency rates.

Figure C2. Compared to unstabilized data, stabilized data showed a weaker relationship between variability and sample size for the annual proficiency rates in the Targeted Support and Improvement model



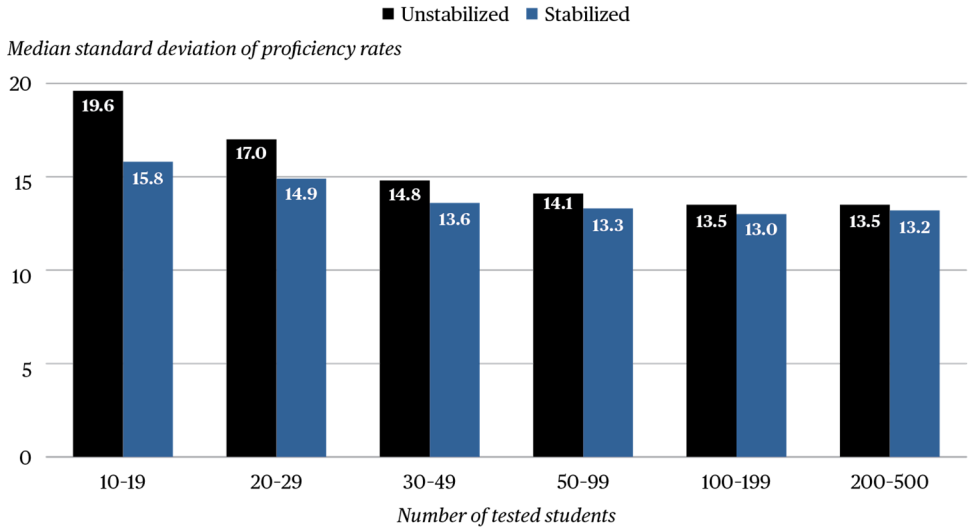
Note: The rows present results for three of the eight student subgroups included in the study. In each panel, the horizontal axis represents the number of tested students in a school-subgroup combination, while the vertical axis represents the academic proficiency rate in that school-subgroup combination. Each data point represents the annual average proficiency rate for a given school and subgroup for the 2018/19 academic year.
Source: Pennsylvania Department of Education data.

There is some indication that stabilization moderates the relationship between sample size and variability for TSI, but it is less pronounced than for ATSI, likely because the annual proficiency rate estimates used in PDE's TSI accountability rules are inherently less stable than the two-year averages used in the ATSI accountability rules.

To provide quantitative support for these relationships, the study team calculated the standard deviation of unstabilized and stabilized proficiency rates across schools separately for each subgroup in each of six categories defined by the number of tested students in the subgroup: 10-19, 20-29, 30-49, 50-99, 100-199, and 200-500.

Figure C3 shows the relationship between sample size and the standard deviation of proficiency rates separately for unstabilized and stabilized estimates.

Figure C3. Stabilization substantially reduced the variability of proficiency rates for small subgroups in the Targeted Support and Improvement model, making the median standard deviation relatively constant across sample size categories



Note: The figure shows the median across subgroups of the standard deviation calculated across schools of a certain sample size within a subgroup for the 2018/19 year. For example, included in the left-most bar are the standard deviation of proficiency rates across subgroups of economically disadvantaged students with only 10-19 test-takers, along with the standard deviations of proficiency rates in other subgroups with only 10-19 test-takers, separately by subgroup. The median standard deviation across subgroups for the 10-19 sample size category is plotted in the figure.
Source: Pennsylvania Department of Education data.

As in figure C2, there is a marked decrease in the standard deviation of proficiency rates with increasing sample size (dark bars). However, standard deviations for both unstabilized and stabilized proficiency rates are higher for the TSI model than for the ATSI model (see figure 3 in main report), and the TSI model does not achieve the same level of similarity in stabilized standard deviations across sample size categories. As noted, annual proficiency rates are slightly less stable than two-year averages, so larger standard deviations and less stabilization are expected for the TSI model. However, even for the TSI model, stabilization cuts the range of standard deviations—the difference between the largest and the smallest standard deviation—roughly in half, from 6.1 percentage points to 2.5 percentage points. Moreover, the median standard deviation of stabilized results for subgroups of 10-19 students (median 15.8, interquartile range [IQR] 14.6-16.9) is *lower* than the median standard deviation of unstabilized results for subgroups of 20-29 students (median 17.0, IQR 14.6-20.1). These results suggest that stabilization could allow smaller subgroups to be included in accountability calculations without increasing the risk of erroneously identifying schools based on measurement error.

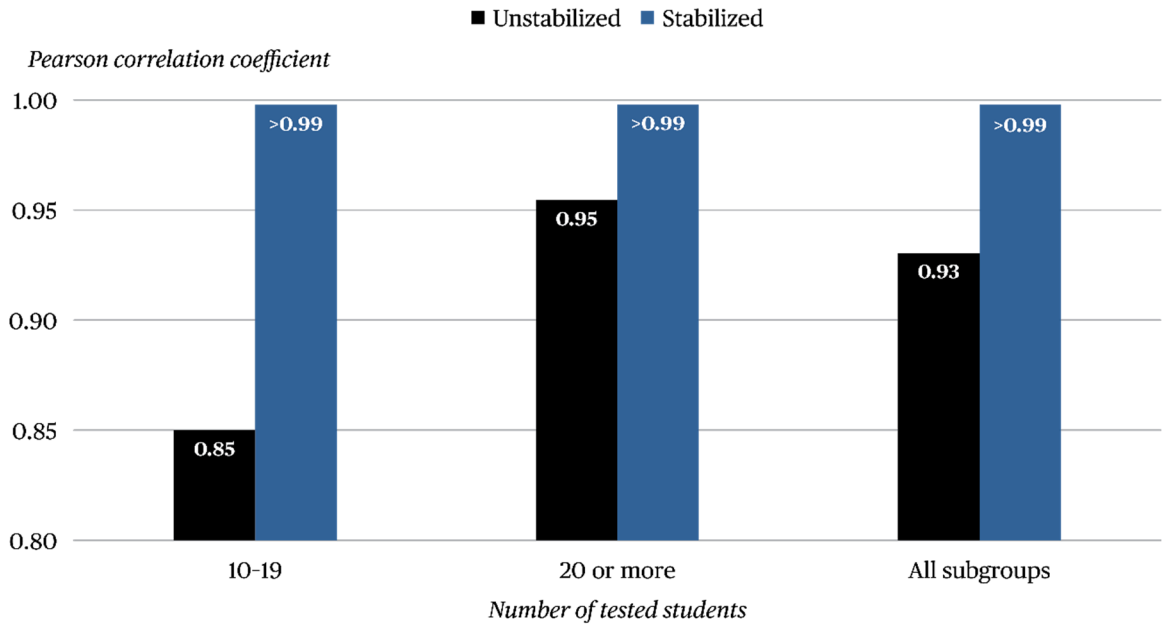
Stabilization improved the reliability of annual proficiency rates, as approximated using year-on-year correlations in proficiency rates

As an additional investigation of the effect of stabilization on the statistical reliability of proficiency rates, the study team calculated the Pearson correlation between the proficiency rates in consecutive academic years: 2015/16 and 2016/17, 2016/17 and 2017/18, and 2017/18 and 2018/19. This correlation approximates statistical reliability insofar as it assesses the stability of proficiency rates within a school-subgroup combination over time. However, this metric is a proxy for preferred reliability measures, such as the split-half sample correlation, which cannot be calculated without student-level data. When interpreting the results, it is also important to note that

heterogeneous error inflates the Pearson correlation coefficient (Lahiri & Suntornchost, 2015), implying that estimated correlations may be overstated for both the stabilized and unstabilized estimates. The study team performed this calculation using unstabilized proficiency rates from each of the three years, as well as proficiency rates stabilized using the TSI (longitudinal) model. This calculation was not performed using two-year averages, as in the ATSI (cross-sectional) model, because the ATSI model produces only one time period of stabilized proficiency rates (average across 2016/17 and 2017/18 years).

The calculated correlations suggest that the statistical reliability of unstabilized proficiency rates is reasonably high and improves with stabilization, especially for subgroups with fewer than 20 tested students. Figure C4 shows the correlations calculated for the most recent pair of years (2017/18 and 2018/19) for stabilized and unstabilized proficiency rates in each of three categories: subgroups with 10–19 students, subgroups with 20 or more students, and overall across all subgroups. Results for other pairs of years were similar.

Figure C4. Stabilization increased the correlation between 2017/18 and 2018/19 proficiency rates in the Targeted Support and Improvement model, especially for subgroups with 10–19 tested students



Note: The horizontal axis specifies the subset of subgroups included in the calculation (whether restricted to those with a certain range of sample sizes or overall), and the vertical axis gives the estimated Pearson correlation coefficient. The horizontal axis begins at a correlation of 0.80 to show that stabilized estimates do not achieve a perfect correlation of 1.0.
Source: Pennsylvania Department of Education data.

For unstabilized estimates, year-on-year correlations were markedly lower for subgroups with 10–19 tested students, indicating that annual proficiency rates were less statistically reliable for these subgroups than for larger subgroups. However, year-on-year correlations were very stable at close to 1.0 for stabilized proficiency rates, regardless of the number of tested students in the subgroup. This finding corroborates the evidence in the previous section showing that stabilization moderates the relationship between sample size and variability, indicating that stabilization also moderates the relationship between sample size and statistical reliability. However, the Pearson correlation is at best a rough estimate, and likely an exaggerated one, of statistical reliability, so these findings are merely suggestive of the potential improvements in reliability with stabilization.

Reference

Lahiri. P., & Suntornchost, J. (2015). Variable selection for linear mixed models with applications in small area estimation. *Sankhya B, 77*(2), 312–320.