IES Institute of Education Sciences

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

Regional Educational Laboratory Mid-Atlantic

At Mathematica

REL 2023-001
U.S. DEPARTMENT OF EDUCATION

A Publication of the National Center for Education Evaluation and Regional Assistance at IES

This report was prepared for IES under Contract 91990022C0012 by Mathematica, Inc. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

February 2023

Forrow, L., Starling, J., and Gill, B. (2023). *Stabilizing subgroup proficiency results to improve the identification of low-performing schools* (REL 2023-001). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/rel/Products/Publication/106926.

This report is available on the Institute of Education Sciences website at https://ies.ed.gov/ncee/rel/.

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

Lauren Forrow, Jennifer Starling, and Brian Gill

February 2023

The Every Student Succeeds Act requires states to identify schools with low-performing student subgroups for Targeted Support and Improvement or Additional Targeted Support and Improvement. Random differences between students' true abilities and their test scores, also called measurement error, reduce the statistical reliability of the performance measures used to identify schools for these categorizations. Measurement error introduces a risk that the identified schools are unlucky rather than truly low performing. Using data provided by the Pennsylvania Department of Education, the study team used Bayesian hierarchical modeling to improve the reliability of subgroup proficiency measures and demonstrate the approach's efficacy.

Why this study?

The Every Student Succeeds Act requires states to identify their lowest-performing schools for Comprehensive Support and Improvement and to identify schools with low-performing student subgroups for Targeted Support and Improvement (TSI) or Additional Targeted Support and Improvement (ATSI). Identifying the schools that most need support hinges on accountability data that reliably measure school performance, as defined by the state. However, random differences between students' true abilities and their test scores–formally called measurement error–can obscure a school's true performance. This is especially likely in small schools or student subgroups, where random factors that affect a small number of students can have an outsized impact on the school's or subgroup's average score. As a result, accountability data for small schools and subgroups might not reliably reflect their performance: small schools and subgroups could be identified for additional support primarily because of bad luck. States typically seek to reduce the risk of these errors by setting a minimum number of students for a subgroup to be included in accountability measures. However, a minimum subgroup size requirement might not fully solve the problem because data for the smallest included subgroups are still likely to be less reliable than data for larger subgroups. At the same time, such minimum subgroup size requirements come at the cost of blocking some subgroups from contributing to a school's performance measures.

For this study, the Mid-Atlantic Regional Education Laboratory (REL Mid-Atlantic) investigated the potential of Bayesian hierarchical modeling to increase the robustness to measurement error of the Pennsylvania Department of Education's (PDE's) accountability data, thereby improving accuracy. Specifically, the study team implemented a Bayesian modeling approach to stabilize the subgroup proficiency rates used to identify Pennsylvania schools for TSI and ATSI. This modeling approach is referred to as Bayesian stabilization, or stabilization for short, throughout the report. (See appendix A for a brief review of the literature on stabilization.)

Proficiency rates are just one of six accountability indicators in Pennsylvania's accountability system.[1] Although academic proficiency by no means provides a comprehensive picture of school performance, it is well suited to this pilot study

For additional information, including a literature review, data and methods, and supplemental results, access the report appendixes at https://ies.ed.gov/ncee/rel/Products/Publication/106926.

1. The six indicators are academic proficiency, academic growth, progress toward fluency for English learner students, career readiness, regular attendance, and graduation rates.

because historical data are readily available and because there is a clear mechanism through which measurement error could affect a school's observed proficiency rate. Acknowledging that academic proficiency is just one dimension of school performance and could reflect factors outside the school's control, such as parental involvement or motivation, the REL Mid-Atlantic team considers this study a proof of concept that could inform Pennsylvania's or other states' decisions to apply stabilization to proficiency or other accountability indicators.

The study goal was to determine whether Bayesian stabilization improves the reliability of the academic proficiency rates used in PDE's accountability system, focusing on TSI and ATSI identification. TSI and ATSI walk a tightrope between reliability and inclusivity. On the one hand, performance data for small subgroups may largely reflect measurement error and should therefore not be used for accountability calculations, lest schools be identified for additional support based on bad luck. On the other hand, excluding small subgroups from accountability calculations for fear of assessing them based on unreliable data may bar them from receiving the support they need. A Bayesian stabilization approach could alleviate this tension by improving the precision and plausibility of proficiency rates for small subgroups, perhaps even for subgroups that are smaller than current minimum sample sizes used for accountability calculations.

This approach could both improve the reliability of TSI and ATSI identifications and expand the set of schools and subgroups included in the accountability system, allowing PDE to target the schools and students that most need additional support. Specifically, informed by the results of this study, the REL Mid-Atlantic has collaborated with PDE to incorporate stabilization as a safe harbor alternative in its 2022 ATSI calculations, so that stabilization can move schools out of, but not into, ATSI. This approach can address both underlying measurement error in accountability indicators and the implications of accountability data collection disruptions in 2020 and 2021 due to the COVID-19 pandemic. The study's findings could be relevant to states across the country, which all face the same need to identify schools for TSI and ATSI, and could address the same tension between an inclusive approach to subgroups and measurement error when small numbers of students are involved.

Research questions

The study explored the usefulness of stabilization for school accountability calculations through the following research questions:

1. How much do stabilized proficiency rates differ from unstabilized proficiency rates? Do these differences vary with the number of students in the subgroup?

2. How much does the Bayesian stabilization approach increase the reliability (long-run stability) of proficiency rates? For small subgroups, are these improvements sufficient to reduce the minimum threshold for a subgroup size from 20 students to 10?

3. How does the set of schools identified as eligible for ASTI change when stabilized proficiency rates rather than unstabilized proficiency rates are used in the identification process?

Research question 1 gauged the impact of stabilization on proficiency rates, overall and by the number of tested students. This analysis determined whether, as context for further analysis, stabilization made a difference to schools' estimated proficiency rates. At the same time, it served as a check that the relationships between unstabilized and stabilized proficiency rates followed the pattern to be expected from theory and the literature, namely that stabilization would have a larger effect on smaller subgroups.

Research question 2 investigated how much stabilization increased statistical reliability and specifically how much stabilization improved reliability for small subgroups that are currently excluded from PDE's

accountability system because they do not meet the minimum sample size requirement. The minimum sample size requirement is intended both to preserve student privacy and to guarantee a minimum level of statistical reliability so that the accountability system focuses on schools and subgroups where enough information is available to assess performance. Pennsylvania currently sets the minimum sample size at 20 students. If stabilization achieves a comparable level of statistical reliability for subgroups with 10–19 students as achieved without stabilization for subgroups with 20 or more students, the benefits of including small schools and subgroups in accountability calculations might outweigh the statistical reasons to exclude them. Specifically, the current minimum subgroup size requirements eliminate small schools and subgroups from consideration for additional support that they might need. Including smaller schools and subgroups in accountability calculations would make them eligible to receive this support.

Research question 3 explored how the effects of stabilizing proficiency rates affected which schools were identified for ATSI.[2] This analysis assessed the possible impact of stabilization on accountability calculations, not its effect on the reliability of accountability data. Although it is not possible to assess the accuracy of accountability decisions, if stabilization improves the reliability of accountability measures, it is also likely to increase the accuracy of ATSI and TSI identifications.

See box 1 for a summary of data sources, sample, methods, and limitations and appendix B for technical details.

Box 1. Data sources, sample, methods, and limitations

Data and sample. The Pennsylvania Department of Education (PDE) provided school-level data for this study. The dataset contains one record for each combination of school and student subgroup for each school year from 2015/16 through 2018/19, for all schools—elementary, middle, and high schools—included in accountability calculations in those years (2,678 schools). For each school, the analysis included all tested students across courses and grade levels, in the following eight subgroups: racial/ethnic categories (Asian students, Black students, Hispanic students, White students, multiracial students); economically disadvantaged students; students with disabilities; and English learner students. The Native American/Alaska Native and Hawaiian/Pacific Islander student groups were excluded because there were fewer than five schools with at least 10 students in those subgroups. Key variables included the percentage of students scoring at or above the state's threshold for academic proficiency in each school-subgroup combination and the number of tested students in each school-subgroup combination in each year. For accountability purposes, Pennsylvania uses an average proficiency rate across math and English language arts (ELA); throughout the report, this average of math and ELA proficiency rates is called the school's proficiency rate.

Methods. The study team implemented two Bayesian statistical models to stabilize PDE's academic proficiency rates. For model statements and technical details, see appendix B.

Additional Targeted Support and Improvement (ATSI) model: The first model mirrors PDE's ATSI identification process. This model stabilized average student proficiency rates across two academic years, 2016/17 and 2017/18. The model is cross-sectional, with one data point (the average proficiency rate across two years) per school-subgroup combination. Results for the ATSI model appear in the main report.

Targeted Support and Improvement (TSI) model: The second model mirrors PDE's TSI identification process. This model stabilized annual student proficiency rates across four academic years, 2015/16, 2016/17, 2017/18, and 2018/19. The model is longitudinal, with one data point per school-subgroup-year combination. To avoid redundancy with ATSI results, results for the TSI model appear in appendix C.

After fitting the models, the study team assessed the effect of stabilization on statistical reliability. In the absence of an error-free benchmark or student-level data with which to compute conventional reliability metrics, the study team

2. The study 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.

approximated reliability by comparing the relationship between sample size and variation in proficiency rates for unstabilized and stabilized estimates. In less reliable estimates, small sample size is associated with more variation–for example, a wider range or larger standard deviation–whereas in more reliable estimates, the correlation between sample size and variation will be weaker. To gauge the effect of stabilization on ATSI identifications, the study team compared unstabilized and stabilized proficiency rates for the set of schools and subgroups identified for ATSI using PDE's accountability rules.

Limitations. The primary limitation of this study is that it was not possible within the study timeframe to obtain access to student-level data. Using school-level data constrained the set of models included in the analysis, which could understate the benefits of stabilization. Without student-level data, it is not feasible to calculate classical measures of statistical reliability, so the study team relied on visualizations and descriptive analysis to assess the extent to which stabilization improves the reliability of academic proficiency rates.

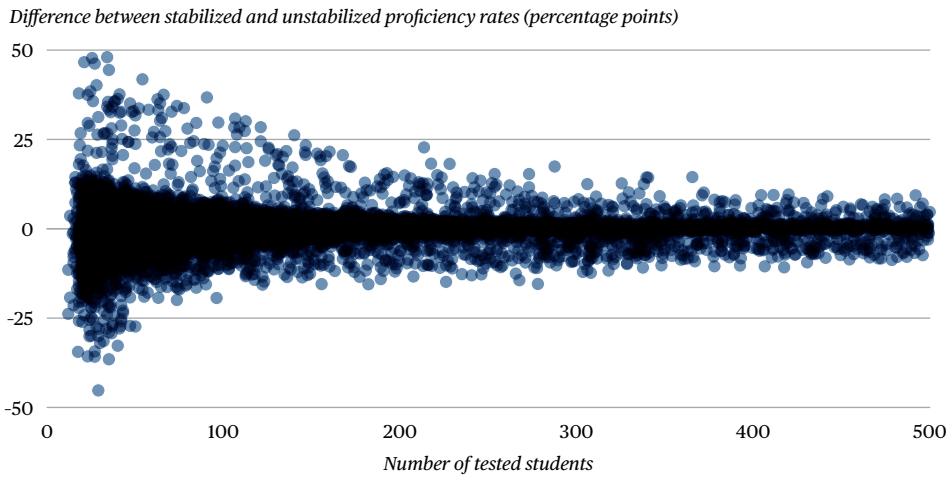
Findings

Stabilization had the greatest effect on small school-subgroup combinations

Average stabilization–the difference between stabilized and unstabilized estimates–for ATSI subgroup proficiency was 3.5 percentage points in each direction; average stabilization was less than 4.5 percentage points for 75 percent of subgroups. Results for TSI were similar (see appendix C).

Stabilized proficiency rates differed from unstabilized rates particularly for small school-subgroup combinations. Figure 1 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. Large school-subgroup combinations were minimally affected.

Figure 1. Stabilization was more influential for smaller subgroups (fewer than 100 tested students) than larger ones for the Additional Targeted Support and Improvement model



Note: Each data point represents the two-year average of academic proficiency rates for a given school and subgroup for the combined 2016/17 and 2017/18 academic years. The horizontal axis represents the number of tested students in that school-subgroup combination, and the vertical axis represents the difference between the stabilized and unstabilized proficiency rates for that school-subgroup combination. 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.

Stabilization moderated the relationship between subgroup size and variation in proficiency rates, indicating an improvement in statistical reliability that could permit lower minimum sample sizes

The unstabilized data also showed a characteristic funnel pattern: proficiency rates varied more among smaller school-subgroup combinations than among larger school-subgroup combinations. There is little reason to expect that variation in true academic proficiency would be greater for small groups of students than for large groups of students, so this relationship more likely reflects measurement error—instability in the proficiency rate estimates for smaller school-subgroup combinations, due simply to their size.

To assess how much stabilization increased statistical reliability, the study team visually compared the relationship between variability and sample size in unstabilized and stabilized proficiency rates. For three of the student subgroups included in the study, figure 2 depicts the relationship between the number of tested students and the proficiency rate for the stabilization model, mirroring data used in PDE's ATSI accountability rules. The panels show this relationship for unstabilized proficiency rates (left column) and stabilized proficiency rates (right column). In all rows, the unstabilized points display the characteristic funnel shape, spanning a wider range on the vertical axis for smaller sample sizes than for larger sample sizes, with smaller subgroups much more likely to have extreme values due to measurement error. As a result of a greater susceptibility to measurement error, smaller subgroups are more likely than larger subgroups to trigger ATSI identification just by bad luck.

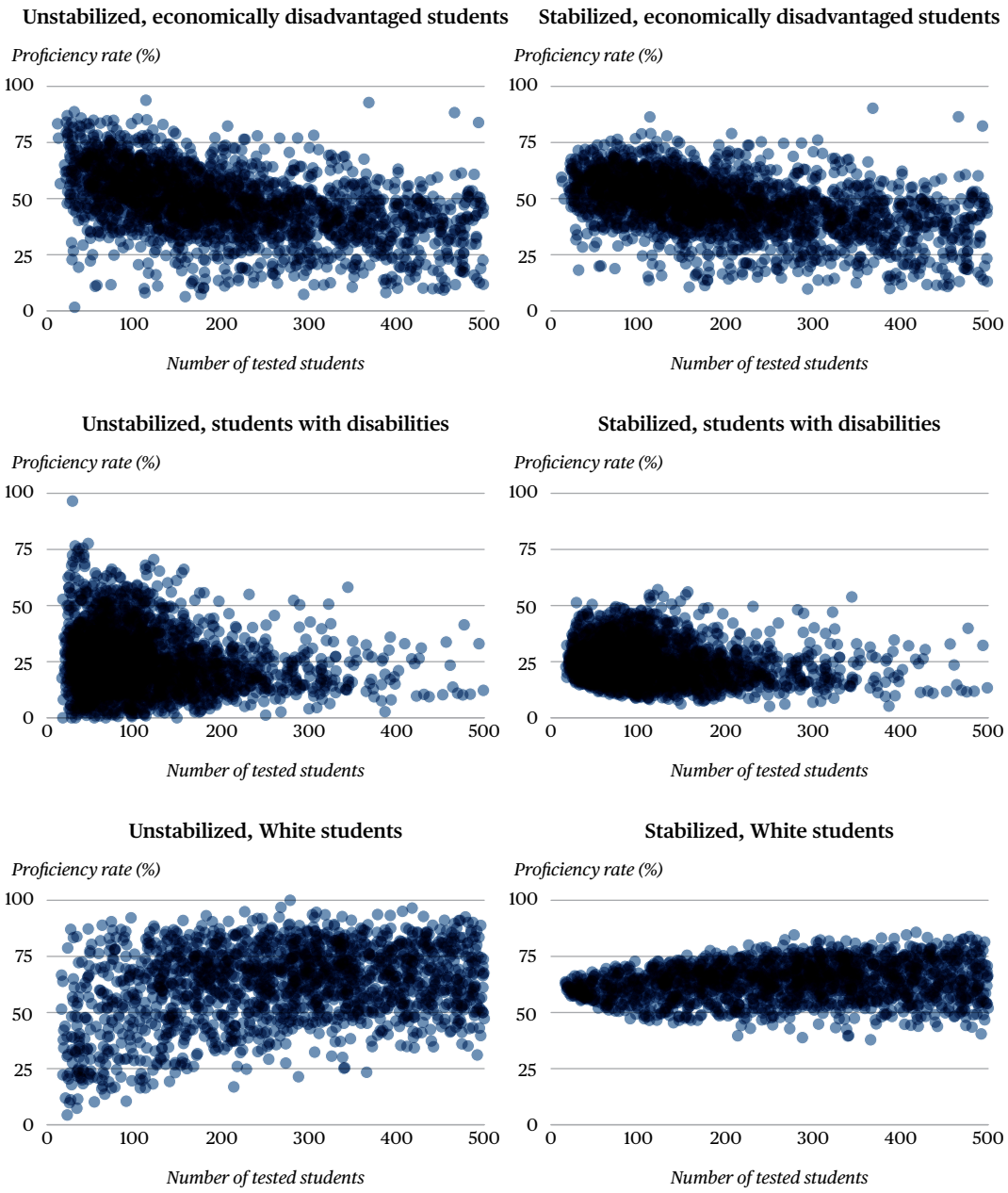
In the stabilized points (right column) in figure 2, by contrast, a relationship between subgroup size and the range of estimated proficiency is less evident. Increased consistency in the estimates' variability across subgroup sizes suggests that stabilization has improved the statistical reliability of the proficiency rates.

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 sample size categories defined by the number of tested students in the subgroup: 10–19, 20–29, 30–49, 50–99, 100–199, and 200–500 tested students. Within a sample size category, the study team then took the median and interquartile range (IQR) of the subgroup-specific standard deviations as a summary of the distribution across subgroups. Finally, these median standard deviations of unstabilized and stabilized proficiency rates were compared by sample size category (figure 3).

The medians of subgroup-specific standard deviations of unstabilized proficiency rates decrease with increasing sample size (dark bars). This relationship reaffirms the pattern in figure 2, where proficiency rates were more widely dispersed for smaller subgroups than for larger subgroups. By contrast, the medians for stabilized proficiency rates are more similar across subgroups (light bars). Because the additional variation in smaller subgroups is likely to reflect measurement error, this result indicates that the stabilized rates are more reliable —less prone to measurement error—than the unstabilized rates. Indeed, the consistency of the median standard deviations of stabilized proficiency rates across sample size categories suggests that the stabilized proficiency rates of smaller subgroups (10–19 students; median 10.4, IQR 5.5–12.4) are more reliable than the unstabilized proficiency rates of larger subgroups (20 or more students; median 15.8, IQR 14.0–18.3).[3] Thus, stabilization may make it possible to include smaller subgroups in accountability calculations without sacrificing statistical reliability.

3. The wider IQR for stabilized than unstabilized standard deviations reflects variability in the amount of stabilization across subgroups.

Figure 2. Compared to unstabilized data, stabilized data showed a weaker relationship between variability and sample size for the Additional Targeted Support and Improvement model



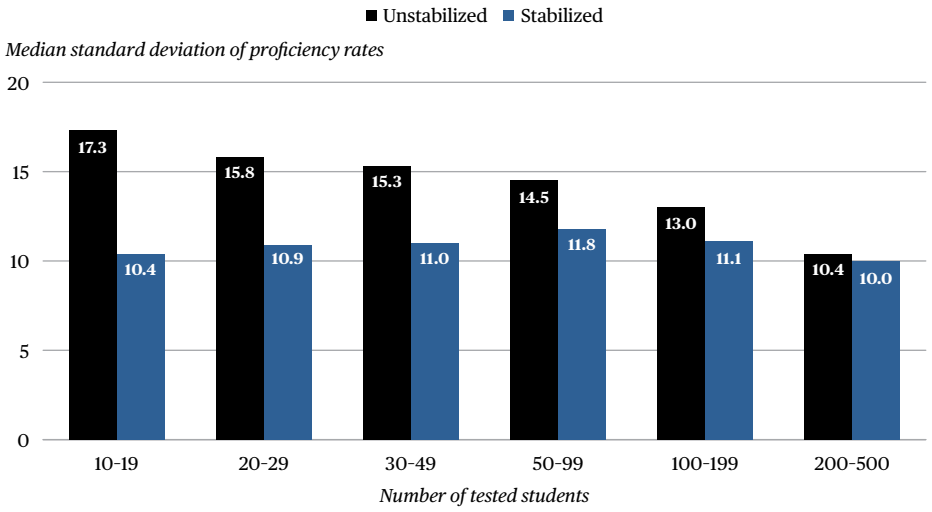
Note: The rows present results for three of the eight student subgroups included in the study. Each data point represents the two-year average of academic proficiency rates for the given school and subgroup for the combined 2016/17 and 2017/18 academic years. In each panel, the horizontal axis represents the number of tested students in that school-subgroup combination, and the vertical axis represents the difference between the stabilized and unstabilized proficiency rates for that school-subgroup combination.

Source: Pennsylvania Department of Education data.

Stabilization moved a subgroup above the proficiency cutoff for Additional Targeted Support and Improvement (ATSI) for 9 of 193 schools identified as ATSI by the Pennsylvania Department of Education, enabling resources to be directed to the schools most likely to truly need support

After concluding that stabilization increased the reliability of academic proficiency rates, the study team assessed the effect of replacing unstabilized proficiency rates with stabilized proficiency rates using PDE's ATSI

Figure 3. Stabilization substantially reduced the variability of proficiency rates for small subgroups in the Additional Targeted Support and Improvement model, making the median standard deviation relatively constant across sample size categories



Note: Data are two-year averages of academic proficiency rates for the combined 2016/17 and 2017/18 academic years. Sample size categories are on the horizontal axis, and the medians of subgroup-specific standard deviations in proficiency rates are on the vertical axis. Darker bars show the median of subgroup-specific standard deviations of unstabilized estimates for a certain sample size category, while lighter bars show the median subgroup-specific standard deviation of stabilized estimates for that sample size category.

Source: Pennsylvania Department of Education data.

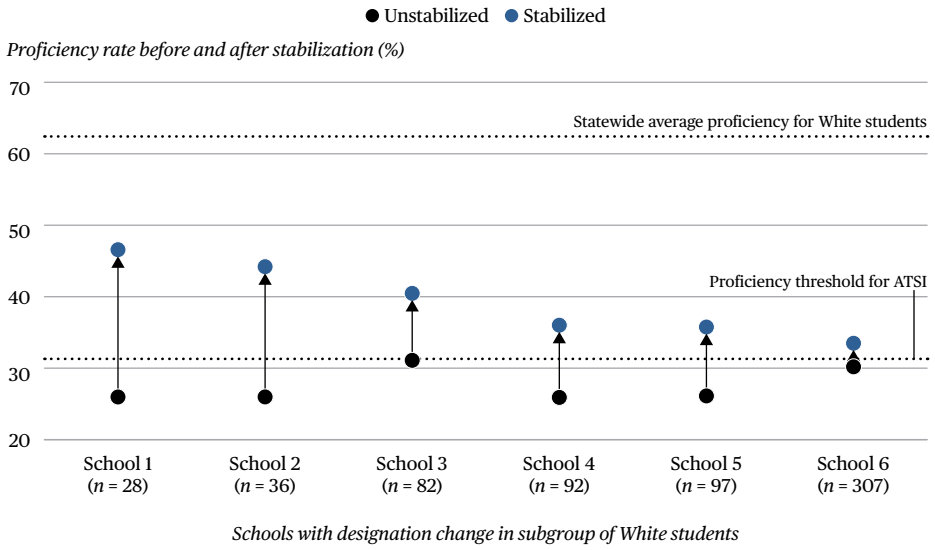
accountability rules. PDE applies a two-phase system of decision rules to identify schools for ATSI. In the first phase, PDE assesses whether schools are eligible for ATSI based on the combination of their academic proficiency rate and academic growth rate relative to performance cutoffs that vary by the school's academic proficiency rate. For example, a school with a proficiency rate between 20 and 30 percent must receive a negative academic growth rate to be eligible for ATSI, whereas a school with a proficiency rate between 10 and 19.9 percent is eligible for ATSI if its academic growth rate is below a designated positive threshold. In the second phase, PDE considers the eligible schools' performance on the remaining accountability indicators relative to performance cutoffs for those indicators, which are determined each year based on the distributions of the observed data.

Using the same calculation, the study team compared ATSI identifications computed using PDE's accountability rules with unstabilized rates with identifications computed using stabilized rates; the accountability data for the five remaining accountability indicators were not adjusted in any way.[4] Stabilization did not substantially change ATSI identifications. Of the 2,678 schools in the analytic dataset (90 of which were identified for Comprehensive Support and Improvement and so were not eligible for ATSI identification), 193 were identified as ATSI by PDE. Of these 193 schools, 9 had a subgroup that switched from ATSI to non-ATSI status under stabilization. In these nine schools, that switch did not result in a change to the school's overall ATSI identification because in each school at least one other subgroup retained its ATSI status after stabilization. In other words, even though stabilized results did not remove any school's ATSI status, stabilization provided information allowing the schools to better focus on the subgroups for which true performance was most likely in the ATSI range.

The nine subgroup changes from ATSI to non-ATSI were observed among the White subgroup in six schools (figure 4) and among economically disadvantaged students in three schools (not shown). Stabilization would be expected to move the White and economically disadvantaged subgroups from ATSI to non-ATSI (rather than in

4. In practice, PDE may use its discretion to adjudicate difficult cases. For this comparison, the accountability rules were applied with no discretionary adjustments.

Figure 4. Stabilization changed the Additional Targeted Support and Improvement (ATSI) identification for the subgroup of White students in 6 of the 193 schools identified as ATSI by the Pennsylvania Department of Education, pulling their proficiency rates above the cutoff



Note: Data are averages for the combined 2016/17 and 2017/18 academic years.

Source: Pennsylvania Department of Education data.

the opposite direction) because stabilization pulls individual subgroups' proficiency rates toward the mean proficiency rate in that subgroup,[5] and in these subgroups the mean proficiency rate is above the ATSI proficiency cutoff.

In principle, stabilization could also reduce a subgroup score enough to move a school into ATSI status—for example, if the mean proficiency rate in the subgroup were below the proficiency cutoff. The study team did not examine that possibility here, for two reasons. First, this is not likely to occur frequently because statewide average proficiency rates are usually higher than the ATSI proficiency cutoff; low-scoring schools are therefore more likely to see stabilized scores move up than down. (The only subgroups in Pennsylvania with average proficiency rates below the ATSI proficiency cutoff are students with disabilities and English learner students.) Second, Pennsylvania is using stabilized scores only as a safe harbor that could remove a school from ATSI status but not move a school into ATSI status.

Limitations

This study's results suggest that stabilization has the potential to improve the statistical reliability of school accountability measures. However, the study's use of school-level rather than student-level data limits both the methods employed in the study and the conclusions that can be drawn from the results.

The use of school-level data constrains the set of models that the study team tested and the metrics used to evaluate them. Without student-level data, the study team could not take advantage of repeated measures of students over time or of student-level variability. To avoid overstating the statistical precision of the results due

5. The core assumption of the Bayesian stabilization models implemented in this study is that, absent information to the contrary, a school's proficiency rate will be close to the overall average proficiency rate. Because the study team fit separate stabilization models for each subgroup, the assumption becomes that a specific school-subgroup's proficiency rate is likely to be close to the average proficiency rate for that subgroup across schools. As a result, the proficiency rates for smaller school-subgroup combinations are pulled toward their subgroup mean.

to students who belong to more than one subgroup, the study team fit models separately for each subgroup, rather than capitalizing on relationships across subgroups. Nonetheless, both the ATSI and TSI models overstated precision to some degree, even within a subgroup, because it was not possible to adjust appropriately for correlation across students whose test scores were included in more than one academic year.

In addition, without student-level data, the study team could not calculate traditional measures of statistical reliability. Instead, the study team relied on visualizations and other descriptive assessments to gauge whether stabilization achieved the desired goals.

Finally, most of the subgroups with accountability status changes had proficiency rates close to the ATSI cutoff, with the stabilized proficiency just on the other side of the cutoff. For these borderline cases, binary decision rules for ATSI identification may overstate confidence in the available evidence about school or subgroup performance. Accounting for statistical uncertainty could allow for a more nuanced understanding. Future work that estimates uncertainty bounds, either around the stabilized proficiencies or around the accountability identifications, could take full advantage of the properties of the Bayesian stabilization models explored in this study.

Implications

The results of this analysis indicate that stabilization could reduce the tension between reliability and inclusivity that characterizes accountability calculations for small schools and subgroups. Stabilization improved the reliability of subgroups' proficiency rates; the relationship between subgroup sample size and variation in proficiency rates, which largely reflects measurement error, was weaker for stabilized than for unstabilized proficiency rates.

Indeed, the results suggest that Bayesian stabilization could permit PDE to reduce its minimum subgroup size from 20 students to 10 while simultaneously reducing the likelihood of erroneously identifying a school for ATSI. In the context of PDE's ATSI calculations, stabilization improved reliability for small subgroups of 10–19 students enough that their distribution of stabilized scores had a similar, or even smaller, variance than the distribution of unstabilized scores of much larger subgroups of up to 200 students (see figure 3). This smaller variance implies that, with stabilization, proficiency rates for smaller subgroups may no longer be too unreliable or imprecise to use in accountability calculations. Including smaller subgroups in accountability calculations could be an important step toward ensuring that all students—even those in small schools and subgroups—receive additional support when they need it. For example, in the current analysis of PDE's historical ATSI calculations, 27 percent of the multiracial subgroups, 18 percent of the Hispanic subgroups, 18 percent of the Asian subgroups, and 17 percent of the Black subgroups had sample sizes of 10–19 students.

Despite increasing the reliability of the estimated proficiency rates, stabilization had little effect on subgroups' identification for ATSI. Nonetheless, stabilized results would provide better information to PDE and to schools about subgroups whose performance is truly below the designated threshold. In about 5 percent of PDE schools identified for ATSI in 2018, stabilization suggested that one subgroup that was below the ATSI proficiency cutoff should have scored above the cutoff. Although these changes did not affect whether the school as a whole would be identified for ATSI, they would permit the school to target improvement efforts more precisely.

On the strength of these results, REL Mid-Atlantic worked with PDE to analyze data through 2022, informing ATSI identifications.

Whether Bayesian stabilization improves the reliability and statistical precision of the estimated proficiency rates is just one of several important considerations for states or districts that contemplate using these

methods. Simplicity and transparency are important qualities of a school accountability system, so that schools understand how they are assessed and deem the identifications credible. Despite better statistical properties, a system that relies on Bayesian stabilization is more complex and potentially more opaque than the existing system. Even so, improvements in the year-on-year consistency of accountability indicator scores, reflecting increased reliability, could gain stakeholders' trust. Individual states and districts must weigh these factors to determine whether the stabilization approach best meets the goals of their accountability system.