IES Institute of Education Sciences

Stabilizing School Performance Indicators in New Jersey to Reduce the Effect of Random Error

Regional Educational Laboratory Mid-Atlantic

At Mathematica

October 2024
REL 2025-009
U.S. DEPARTMENT OF EDUCATION

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

Stabilizing School Performance Indicators in New Jersey to Reduce the Effect of Random Error

Morgan Rosendahl, Brian Gill, and Jennifer E. Starling

October 2024

The Every Student Succeeds Act of 2015 requires states to use a variety of indicators, including standardized tests and attendance records, to designate schools for support and improvement based on schoolwide performance and the performance of groups of students within schools. Schoolwide and group-level performance indicators are also diagnostically relevant for district-level and school-level decisionmaking outside the formal accountability context. Like all measurements, performance indicators are subject to measurement error, with some having more random error than others. Measurement error can have an outsized effect for smaller groups of students, rendering their measured performance unreliable, which can lead to misidentification of groups with the greatest needs. Many states address the reliability problem by excluding from accountability student groups smaller than an established threshold, but this approach sacrifices equity, which requires counting students in all relevant groups.

With the aim of improving reliability, particularly for small groups of students, this study applied a stabilization model called Bayesian hierarchical modeling to group-level data (with groups assigned according to demographic designations) within schools in New Jersey. Stabilization substantially improved the reliability of test-based indicators, including proficiency rates and median student growth percentiles. The stabilization model used in this study was less effective for non-test-based indictors, such as chronic absenteeism and graduation rate, for several reasons related to their statistical properties. When stabilization is applied to the indicators best suited for it (such as proficiency and growth), it leads to substantial changes in the lists of schools designated for support and improvement. These results indicate that, applied correctly, stabilization can increase the reliability of performance indicators for processes using these indicators, simultaneously improving accuracy and equity.

Why this study?

The Every Student Succeeds Act (ESSA) of 2015 requires states to designate their lowest performing schools for Comprehensive Support and Improvement (CSI) and to designate schools with low-performing student groups for Targeted Support and Improvement (TSI) or Additional Targeted Support and Improvement (ATSI).[1] Designating these schools and groups requires data that reliably measure performance on a variety of indicators, including proficiency and growth in reading and math, graduation rates, and other indicators determined by each state.[2] Outside of formal ESSA accountability, reliable indicators of school and student group performance are also diagnostically critical for

For additional information, including a theoretical review, technical methods, supporting analysis, and other analyses, access the report appendices at https://ies.ed.gov/ncee/rel/Products/Publication/108130.

1. This report includes results for simulated CSI and ATSI designations but not for simulated TSI designations. The New Jersey Department of Education identifies schools for TSI based on multiple years of performance data. Analyzing the impact of stabilization on TSI designation requires stabilizing two consecutive years of performance data separately. The model selected for this analysis used all available past data in a single stabilization process, so stabilizing two consecutive years of data is not feasible with the model and data used for this analysis.
2. For ESSA, student groups within a school are defined using demographic designations, including racial/ethnic designations and designations for students with disabilities, economically disadvantaged students, and students who are English learners. This report refers to these designations as *subgroups* and refers to groups of students of a particular designation within each school as *student groups*, sometimes shortened to *groups*. This distinction is useful when discussing scores and group size distributions within designations.

informing school- and district-level decisionmaking about resources and interventions. However, random differences between students' true performance and measurement of their performance—known as measurement error—can undermine the reliability of these indicators. The performance of small groups of students is especially susceptible to measurement error. Small groups are disproportionately likely to have very high or low scores that are due to measurement error rather than to true performance. This means that small groups of students are disproportionately likely to be incorrectly designated as needing support.

States typically seek to reduce the risk of misclassification by setting a minimum number of students for a group in a school to be included in performance indicators. Setting the minimum number, which ranges from 10 to 30 students across states, requires a tradeoff between equity and accuracy. A smaller minimum promotes equity by making student groups visible, but at the cost of reporting unreliable results. A larger minimum reduces the effects of measurement error, making results more reliable at the cost of making many student groups invisible to the accountability system. Neither solution is perfect, and regardless of the minimum group size, indicators for the smallest included groups are likely to be less reliable than indicators for larger groups.

In a recent study, the Regional Educational Laboratory Mid-Atlantic demonstrated the potential of stabilizing performance indicators using Bayesian hierarchical modeling (referred to hereafter as *stabilization*) to reduce measurement error in proficiency rates in reading and math in Pennsylvania (Forrow et al., 2023). The study found substantial improvements in the reliability of school subgroup proficiency indicators, particularly for small student groups.

This study expanded on the Pennsylvania study by applying similar methods to a broader set of performance indicators for New Jersey schools. The New Jersey Department of Education (NJDOE) was interested in having a model that could increase the reliability of multiple performance indicators for diagnostic and accountability purposes. The goal of the model applied in this study was to reduce measurement error and improve the reliability of scores, especially scores from small groups of students. Stabilization considers each group's data in the broader context of data from all groups within the same subgroup and data from the same group in other years and then applies structured assumptions about how these scores relate to each other to reduce the effects of measurement error on performance indicators. This process increases the reliability of performance indicators—especially for small student groups for which measurement error can have an outsized effect—which can simultaneously promote equity and accuracy. The study also examined whether an increase in reliability might create an opportunity to reduce the minimum number of students required to include a student group in accountability.

Research questions

This study explored using stabilization to improve the reliability of school performance indicators. Improved reliability can increase equity by better designating the schools and student groups with the greatest needs and by potentially allowing states to include smaller student groups in their accountability processes. The study sought to address the following three research questions:

1. Does stabilization behave as expected for each performance indicator? That is, for each indicator, does stabilization have a larger effect on scores from smaller student groups, especially when, given the standard deviation, those scores are further from the mean of the score distribution for the subgroup-indicator combination?

2. Does stabilization improve the reliability of performance indicators, especially in groups of 10 to 19 students?

    a. For each indicator, does stabilization have the desired effect of reducing or eliminating the inverse relationship between student group size and score variance?

b. Does stabilization reduce the rates at which groups of 10 to 19 students are overrepresented in the extremes of the score distribution?[3]

3. How does incorporating stabilized performance indicators into CSI and ATSI designations change the set of schools designated as eligible using a designation process implemented without stabilization? This analysis assesses the possible effects of stabilization on accountability decisions rather than its effect on the reliability of underlying accountability data.

Question 1 tests whether the type of stabilization employed in this study is appropriate for each of the performance indicators New Jersey uses, which would allow the state to apply a single stabilization model for all subgroups and indicators. Theory (Stein, 1956; see appendix A for a more detailed theoretical review) and the Pennsylvania study (Forrow et al., 2023) suggest that stabilization would have a larger effect on scores for smaller student groups, for which measurement error also has a larger effect. To verify this suggestion, the study team first checked that stabilization had the expected larger effect on more extreme scores from smaller student groups and examined how stabilization affected various performance indicators differently.

Question 2 investigates the core hypothesis of this study, that stabilized performance indicators should be more reliable than their unstabilized counterparts. Aggregated measurement error may affect CSI and ATSI designations or school- or district-level decisions in ways such that students most in need of support do not receive it. Because increased variance for smaller student groups is reflective of the outsized effect of measurement error, confirming a reduced relationship between group size and score variance is one way to verify that stabilization has improved indicator reliability.

One consequence of the outsized effect of measurement error for small student groups is that very small student groups are often overrepresented in the extremes of score distributions. That can interfere with states' confidence in making CSI and other accountability designations and has prompted states to institute minimum student group sizes for accountability processes. The study examined the extent to which stabilization reduced or eliminated the overrepresentation of groups of 10 to 19 students in the extremes of score distributions for each indicator. If stabilization reduced this effect sufficiently to achieve reliability for student groups of 10 to 19 that is comparable to that achieved for larger student groups without stabilization, New Jersey could consider including these student groups in future designations of accountability in the interest of improving equity in the accountability system.

Question 3 investigates the effect of stabilization on CSI and ATSI designations. Stabilization is performed on individual performance indicators, which are aggregated into composite scores that allow states to make CSI and ATSI designations. Stabilization should affect composite scores consistently with how it affects scores for individual indicators. The study compared designations made using unstabilized and stabilized indicators and compared the number and size of schools and student groups designated using each dataset.

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

3. Currently, New Jersey reports on the performance of student groups as small as 10 but does not include student groups of 10 to 19 in accountability processes due to score instability caused by the outsized effects of measurement error. Stabilizing scores from these small student groups so that they are no longer overrepresented in the extremes of score distributions could allow New Jersey to include them in future accountability processes.

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

Data and sample. The New Jersey Department of Education (NJDOE) provided cleaned and aggregated data for academic years 2015/16 through 2021/22 at the school, year, indicator, and subgroup levels. These data are also available via NJDOE's accountability website (New Jersey Department of Education, 2024). Indicator data varied in availability and in the students to which they applied. Details on which indicators were available in each year are in table B1 in appendix B. Data for school years 2019/20 and 2020/21 were omitted due to teaching and learning disruptions caused by the Covid-19 pandemic.

In keeping with NJDOE's accountability procedures under the Every Student Succeeds Act of 2015, the analysis included schoolwide performance indicators and subgroup indicators for each of the following: racial/ethnic categories, economically disadvantaged students, students with disabilities, and English learner students.

Methods. The study team implemented one model to stabilize each of NJDOE's performance indicators, which was fit to each subgroup-indicator combination independently. After fitting the models, the study team assessed the effect of stabilization on each indicator, overall and by number of students in each group, and checked that stabilization had a larger effect on smaller student groups, as expected. Models were fit using standardized indicators. To facilitate comparison between indicators on different scales, the report presents selected results for standardized unstabilized indicators (unstabilized z-scores) and for their stabilized counterparts (stabilized z-scores, meaning z-scores that have been stabilized rather than stabilized results on a z-score scale, because model outputs are not standardized a second time after stabilization).

To evaluate whether stabilization improved reliability, the study team first compared the relationship between student group size and score variance for each subgroup-indicator combination for unstabilized and stabilized indicators. An inverse relationship between variance and student group size is indicative of a larger effect of measurement error on small student groups, so reducing the relationship between student group size and variance reflects increased reliability for that indicator. Second, the study assessed whether stabilization reduced the overrepresentation of small student groups in the extreme ends of score distributions. Overrepresentation of small student groups in the extremes of the score distribution is also indicative of an outsized effect of measurement error on scores, so reducing that overrepresentation reflects improved reliability and may be sufficient to allow New Jersey to include those student groups in accountability processes.

To gauge the effect of stabilization on CSI and ATSI designations, the study team compared unstabilized and stabilized performance indicators for the set of schools and student groups designated for CSI and ATSI using NJDOE's accountability rules. Because NJDOE designates CSI schools based on relative performance (schools performing below the 5th percentile of all Title I schools in the state), this comparison involved calculating a new threshold for CSI designations from stabilized indicators and assigning new designations based on that threshold.

Limitations. This study required a single model that NJDOE could fit for all performance indicators, producing stabilized scores to inform accountability designations. Such a model must be sufficiently complex to represent data accurately and sufficiently simple to be estimable using available data and software. To achieve this, the model assumes linearity over time, with an adjustment for years after the start of the Covid-19 pandemic. This adjustment protects stabilized performance indicators for pre-Covid years from being affected by unstabilized performance indicators in Covid-affected years, which were often outliers that would otherwise have had an outsized effect on other years.

Modeling choices were also constrained by the use of subgroup-level data rather than student-level data, which were not available. Because subgroup-level data do not capture overlaps between student groups (such as White students who are also economically disadvantaged or economically disadvantaged students with disabilities), each subgroup-indicator combination was stabilized individually, potentially understating the benefits of stabilization were it to be applied to student-level data. Finally, because classical measures of statistical reliability cannot be calculated without student-level data, the study relied on visualizations and descriptive analysis of proxies to assess whether and to what extent stabilization improved the reliability of indicators.

Findings

Stabilization yielded expected adjustments for all indicators, but not all indicators are equally suited to stabilization models that assume normal error distributions

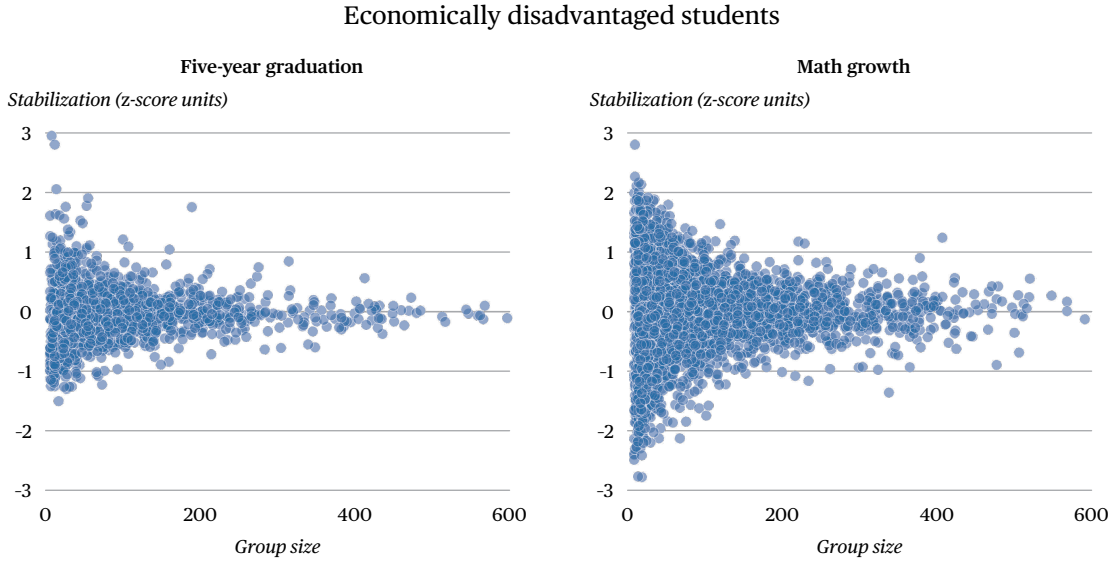
Stabilization accounts for measurement error by making larger adjustments to extreme scores (further from the mean, given the standard deviation of the distribution) and to scores from smaller student groups. To verify that stabilization showed this expected behavior for all subgroup-indicator combinations, data were grouped by subgroup, indicator, and year. Correlation coefficients were calculated between the amount of adjustment and student group size and between the amount of adjustment and the absolute value of unstabilized z-score.

The amount of adjustment was negatively correlated with student group size (average correlation -0.28) and positively correlated with the absolute value of the z-score (average correlation 0.52) for all subgroup, indicator, and year combinations, as expected. The standard pattern of adjustments from a stabilization process is a funnel-shaped distribution with mean near zero and greater variance in adjustment for small student groups (figure 1 illustrates this for two focal indicators in one focal group; see appendix C for more subgroup-indicator combinations). Consistent with expectations, adjustments were generally greater for small student groups, but the extent of the adjustments also depended on their z-score, the variance in the unstabilized distribution, and the group's performance in other years.

By design, stabilization should not radically change most scores, including those for small student groups. To verify this, the study team analyzed the distribution of absolute changes in z-scores. Across all stabilized z-scores, about 82 percent changed by less than 0.5 standard deviation.

All performance indicators are not equally suited to stabilization models that assume a normal distribution. The most common stabilization models, including the model used in this study, assume a normal distribution of data and measurement error and so are not ideally suited to stabilizing data with a distribution that is heavily

Figure 1. Adjustments are small on average, with larger adjustments for smaller groups of students, for which error has a larger impact



Note: Stabilized and unstabilized score data. Data were collected between school years 2015/16 and 2021/22 for five-year high school graduation rates and between 2016/17 and 2018/19 for math growth. Data for school years 2019/20 and 2020/21 were omitted due to disruptions caused by the Covid-19 pandemic.

Source: Analysis of unstabilized score data aggregated and cleaned by the New Jersey Department of Education, also available online (New Jersey Department of Education, 2024).

skewed or has a mean near a boundary.[4] Although stabilization behaved as expected for all subgroups and indicators, not all indicators were equally suited to the stabilization model's assumptions of normal distributions. Additionally, some performance indicators are more complex to measure, have more sources of error, or are more sensitive to error than others and, therefore, stand to benefit more from stabilization.

The study distinguished two types of performance indicators that differ in their suitability for stabilization: test-based and non-test-based. Test-based indicators measure students' academic performance or growth, including math and English language arts (ELA) proficiency and growth and English language proficiency progress. In New Jersey and many other states, non-test-based indicators include chronic absenteeism and four- or five-year high school graduation rates. Chronic absenteeism and graduation rates often take on values at or near their bounds (0 percent or 100 percent), so that score distributions are heavily skewed and poorly aligned with the model's assumption of normal distribution.

Additionally, compared with test-based indicators, attendance and graduation rates are simpler to measure, are measurable directly by educators within a school (as opposed to calculated by a testing company or other contractor), and have fewer sources of measurement error. That is, although these indicators may be affected by transient factors that introduce measurement error—such as a student's chronic illness increasing the absenteeism rate for a small student group in a single year—these factors are relatively few. By contrast, test-based indicators are subject to more sources of error, including how well a student slept the night before the exam, whether they ate breakfast that morning, or whether they were distracted during the exam. Differences in the number and sources of error may produce score distributions for which there is not a strong inverse relationship between student group size and score variance, which can make effective stabilization more challenging. This is because the core premise of stabilization is that measurement error has a larger effect on scores from smaller student groups.

Thus, when choosing whether and how to apply a stabilization model, it is important to understand how well the model aligns with the data. Performance indicators are best suited to a stabilization model when their associated score distribution aligns reasonably with the assumptions of that model. The model selected for this study assumes normally distributed measurement error and scores, meaning that scores are symmetrically distributed about their mean and that scores near the boundaries (0 and 100 for all indicators) are uncommon. Although the selected model allows some departure from normal distributions, indicators might not be well-suited for stabilization under this assumption if their distributions are very far from normal and if they frequently have values at or near the outer bounds of their scales. This is often the case for indicators such as chronic absenteeism and graduation rates.

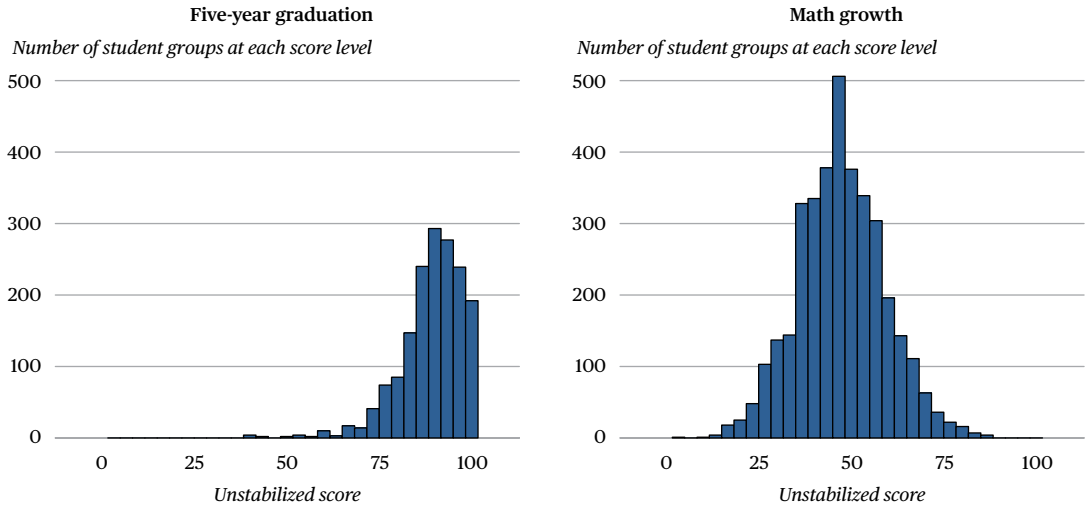
Although there is no universal approach to determining which indicators are well-suited to stabilization models that assume a normal distribution, modelers can perform several simple checks:

- *Histograms of data distributions at the subgroup-indicator level.* Well-suited indicators will have data points that are roughly symmetrically distributed about a single peak, with relatively few data points at the boundaries. Histograms of one indicator that meets this criterion (math growth) and one that does not (five-year high school graduation) are in figure 2, which focuses on economically disadvantaged students. In general, New Jersey's test-based indicators show distributions consistent with model expectations, while its non-test-based indicators do not (see figure C.2 in appendix C).
- *Other checks of model assumptions.* Modelers may combine histograms with other checks to explore how well data distributions align with model assumptions. Other forms of visual inspection, such as Q-Q plots that help check for normality, and scatterplots (like figure 3), can help modelers understand how the data align

4. The assumption that data are normally distributed is also violated in cases where scores tend to saturate near their bounds. There are a variety of methods, of varying complexity, to model scores on bounded intervals. Some are discussed briefly, but it was not within the scope of this study to explore them in detail.

Figure 2. Histograms of the score data for economically disadvantaged students show a heavily skewed distribution for high school graduation rates and a normal distribution for math growth scores

Economically disadvantaged students



Note: Unstabilized score data. Data were collected between school years 2015/16 and 2021/22 for five-year graduation rates and between 2016/17 and 2018/19 for math growth. Data for school years 2019/20 and 2020/21 were omitted due to disruptions caused by the Covid-19 pandemic.

Source: Analysis of unstabilized score data aggregated and cleaned by the New Jersey Department of Education, also available online (New Jersey Department of Education, 2024).

with model assumptions. Statistical tests of normality—such as D'Agostino's K-squared test, and point estimates such as the mean and median of a distribution—offer a concise means to supplement visual inspection. However, because statistical tests are sensitive to sample size, and point estimates provide little information about the shape of a distribution, visual inspection may be most useful for exploring alignment between data and model assumptions, with supplementary numeric tests as desired.

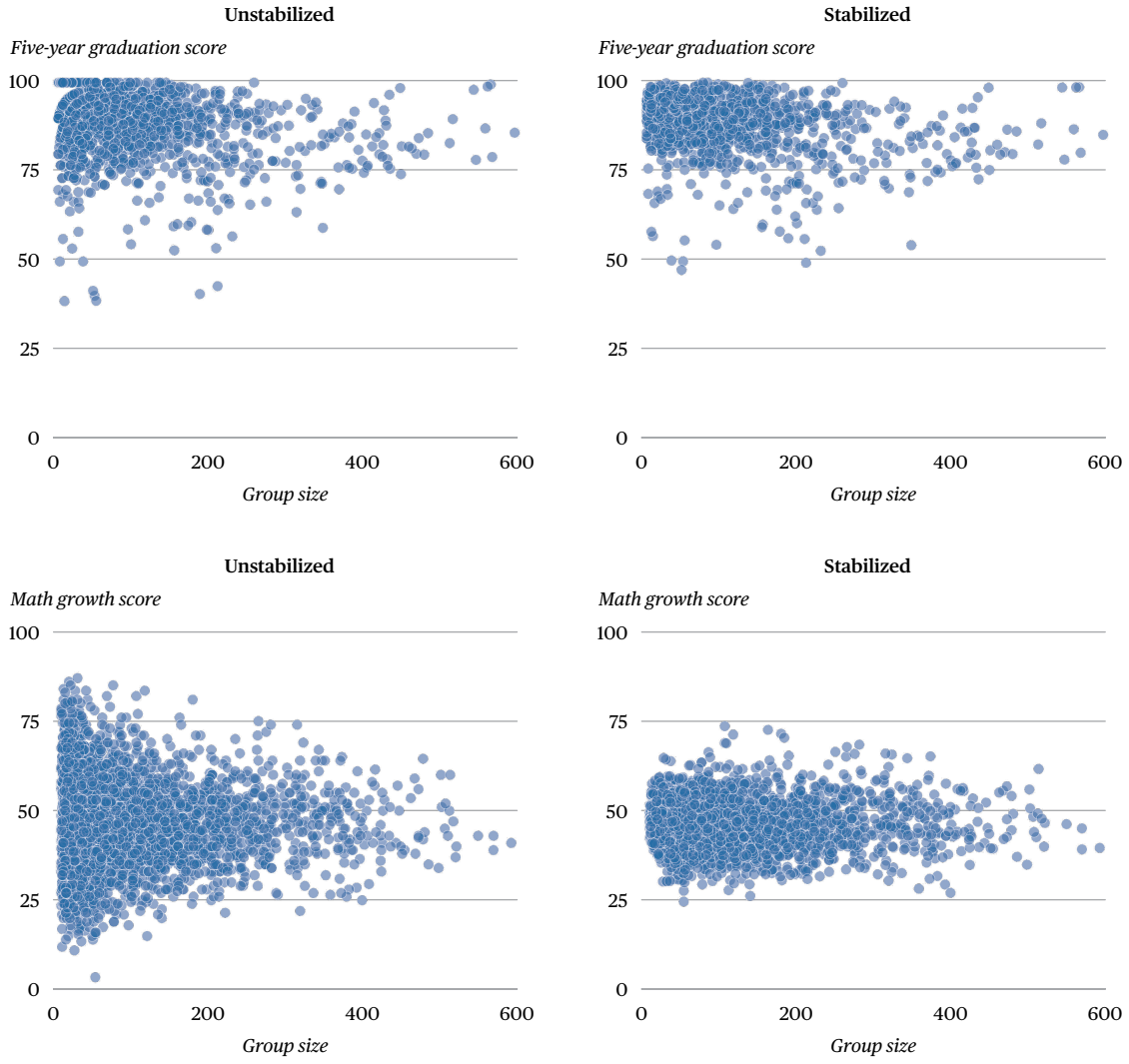
In subgroup-indicator combinations for which model assumptions are very poorly aligned to the data, models may not converge. This is one obvious sign of a mismatch between model and data. However, even in models that appear to converge without issue, stabilized data from indicators that are not well-suited to stabilization by a model will reflect that misalignment.

To illustrate how stabilization affects indicators that are not well-suited to stabilization, the study team compared the effects of stabilization on five-year graduation rates and on math growth. Graduation rates may not be well-suited to stabilization under assumptions of normal distributions because they are often 100 percent, resulting in an asymmetric distribution that saturates at its upper bound; this truncation pushes the mean lower than it would be in an unbounded distribution, potentially leading to excessive downward adjustment of scores at the top of the scale. This report uses the economically disadvantaged students subgroup to examine this issue, but the study found similar patterns across all subgroups. Although values of 100 percent are "extreme" in that they reflect the boundary of the distribution of five-year graduation rates, they were frequent regardless of student group size (see figure 3). Additionally, the unstabilized distributions lacked a clear relationship between graduation rate variance and student group size (figure 4). In short, the unstabilized data did not reflect the model assumptions. In consequence, the model was not equipped to accommodate the plausibility of very high scores and may have imposed excessive adjustments, especially on small student groups.[5]

5. For some subgroup-indicator combinations that were not well suited to stabilization by a model that assumes normal distributions, stabilization changed the shapes of the overall data distributions to be less skewed and more normally distributed. In cases where the unstabilized data are heavily skewed, this reflects the model assumptions overwhelming the data, which can reduce indicator

Figure 3. Unstabilized math growth scores show a clear relationship between group size and score variance that is reduced by stabilization, but five-year graduation scores do not benefit similarly

Economically disadvantaged students



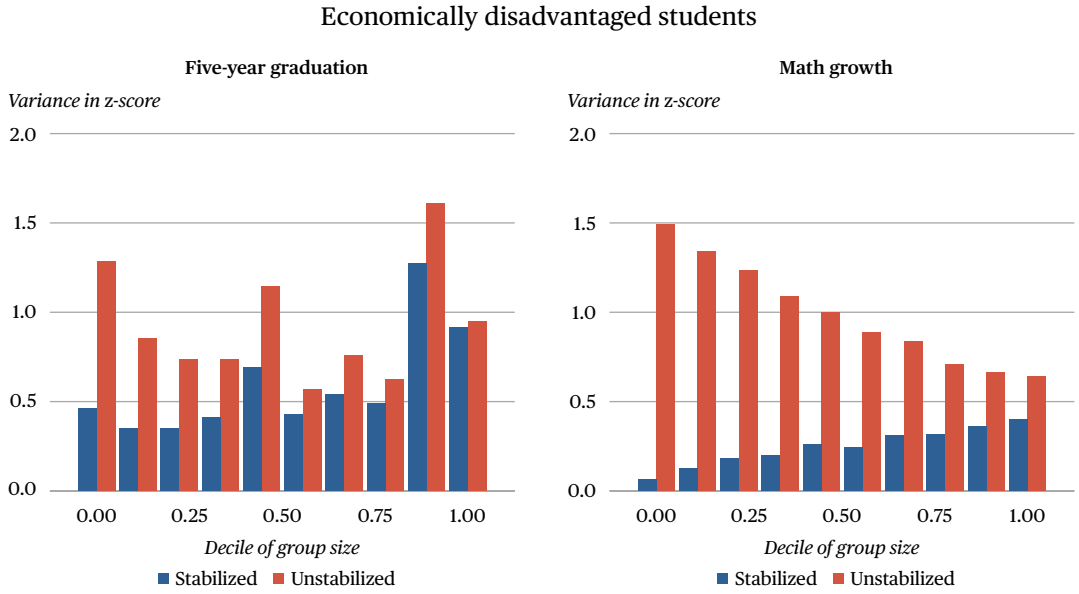
Note: Stabilized and unstabilized score data. Data were collected between school years 2015/16 and 2021/22 for five-year graduation rates and between 2016/17 and 2018/19 for math growth. Data for school years 2019/20 and 2020/21 were omitted due to disruptions caused by the Covid-19 pandemic.

Source: Analysis of unstabilized score data aggregated and cleaned by the New Jersey Department of Education, also available online (New Jersey Department of Education, 2024).

In contrast, math growth should be well suited to stabilization because, by construction of the indicator, extreme scores and heavily skewed distributions are less likely. Although there is no theoretical reason for math growth to be affected by student group size, the distribution of unstabilized scores took on the funnel shape that is the expected result of measurement error, with greater variation in smaller student groups (see figure 3). In addition, unstabilized math growth scores were symmetric around a mean value, with very few scores near the boundaries of 0 or 100 (see figure 3). All of these properties are well-aligned with model assumptions. Stabilization made larger adjustments to more extreme scores from smaller student groups, reducing the relationship between student group size and score variance, as indicated by the reduced funnel shape of the stabilized

reliability rather than improve it. For indicators like these, there are steps that modelers can take to improve suitability, but these are outside the scope of this study.

Figure 4. Stabilization improves the uniformity of variance for math growth but not for five-year high school graduation rates



Note: Stabilized and unstabilized score data. Data were collected between school years 2015/16 and 2021/22 for five-year graduation rates and between 2016/17 and 2018/19 for math growth. Data for school years 2019/20 and 2020/21 were omitted due to disruptions caused by the Covid-19 pandemic.

Source: Analysis of unstabilized score data aggregated and cleaned by the New Jersey Department of Education, also available online (New Jersey Department of Education, 2024).

distributions in figure 3. The reduced relationship between student group size and indicator variance reveals that stabilization, behaving as expected, likely improved the reliability of math growth scores.

Ideally, stabilization should reduce scores' variance somewhat across all student group sizes and increase the uniformity of scores' variance across student group sizes, resulting in a reduced relationship between student group size and variance. For both indicators, stabilization reduced variance across all student group sizes (see figure 4). For math growth, it also increased uniformity of variance across all student group sizes.

Taken together, these results illustrate how stabilization under the assumption of normal distributions is better suited to indicators that are approximately symmetrically distributed and for which scores near the boundaries of the distribution are infrequent. When choosing whether and how to implement stabilization, agencies should thoroughly explore data distributions, supplementing visual checks with numeric ones, to understand how well indicator distributions align with model assumptions of normality. In New Jersey, test-based indicators are better aligned to model assumptions than are indicators of chronic absenteeism and graduation rates. For more details on cases of subgroup-indicator combinations that are less suited to stabilization, see appendix C.

The remainder of this report focuses on the stabilization of test-based indicators only.

Stabilization affected proficiency and growth indicators differently. Both proficiency and growth indicators are well-suited to stabilization in that they are roughly symmetrical, with few scores near the bounds of the distribution. They also show a clear inverse relationship between student group size and score variance that is not theoretically justifiable. That is, there is no reason to believe that smaller groups of students would inherently show greater diversity in their performance on indicators than would larger ones. However, the design of these indicators results in some differences in their distributions and, therefore, in how stabilization affected them.

Stabilization estimates two components of score variance: the true variance and measurement error. True variance is assumed to be approximately constant regardless of student group size, and larger student groups are assumed to have less measurement error, so their variance more heavily informs estimates of true variance. Relative to proficiency, growth has lower score variance in its largest student groups (table 1). In consequence, stabilization will tend to estimate a smaller true variance for growth, and smaller student groups with more extreme z-scores will tend to receive larger adjustments. This results in a larger average adjustment size and a stronger correlation between the absolute value of the *z*-score and adjustment size (see table 1).

To illustrate this difference, *z*-score adjustments are plotted against student group size for growth and proficiency indicators across all student groups (figure 5). Although both indicators show larger adjustments for smaller

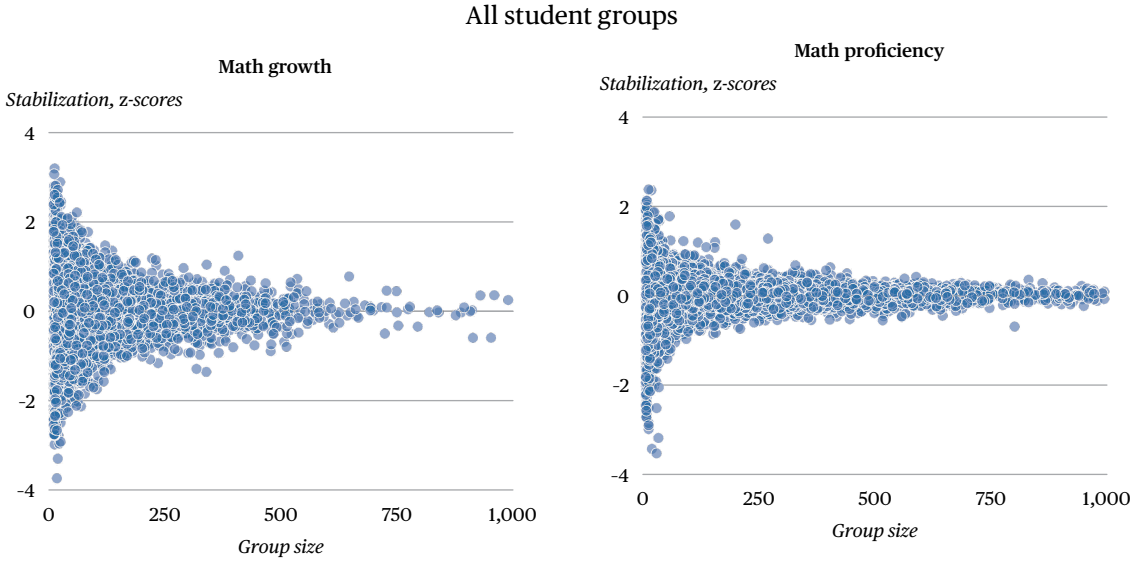
Table 1. Differences in the characteristics of growth and proficiency indicators affect adjustment size

Indicator	Variance of unstabilized *z*-scores of largest 10 percent of student groups	Mean adjustment size (*z*-score scale)	Correlation between adjustment size and absolute value of *z*-score
English language arts			
Growth	0.583	0.51	0.78
Proficiency	0.753	0.25	0.27
Math			
Growth	0.556	0.50	0.77
Proficiency	0.641	0.23	0.26

Note: Stabilized and unstabilized score data. Data were collected between school years 2015/16 and 2021/22 for proficiency rates and between 2016/17 and 2018/19 for growth. Data for school years 2019/20 and 2020/21 were omitted due to disruptions caused by the Covid-19 pandemic.

Source: Analysis of unstabilized score data aggregated and cleaned by the New Jersey Department of Education, also available online (New Jersey Department of Education, 2024).

Figure 5. Growth indicators may show larger adjustments than proficiency indicators at the same student group sizes



Note: Stabilized and unstabilized score data. Data were collected between school years 2015/16 and 2021/22 for proficiency rates and between 2016/17 and 2018/19 for growth. Data for school years 2019/20 and 2020/21 were omitted due to disruptions caused by the Covid-19 pandemic.

Source: Analysis of unstabilized score data aggregated and cleaned by the New Jersey Department of Education, also available online (New Jersey Department of Education, 2024).

student groups, growth indicators show larger adjustments than proficiency indicators for student groups of the same size.

Stabilization improved the reliability of test-based performance indicators, especially for small groups of students

Stabilization's larger effect on scores for smaller student groups is especially pertinent when considering whether to include smaller groups of students (10 to 19 students in New Jersey) in accountability. Outsized measurement error in scores from small student groups produces disproportionately high score variance and, therefore, overrepresentation of small student groups in the extremes of score distributions. Verifying that stabilization mitigates both of these phenomena would suggest that it improves reliability.

First, as there is little reason to believe that true performance would be more variable for small student groups than for large student groups, stabilization should reduce a performance indicator's measurement error and result in similar variation for small and large student groups. To this end, the variance of unstabilized and stabilized *z*-scores was calculated for each test-based indicator among two sets of small student groups—10 to 19 students (not currently included in NJDOE's accountability process) and 20 to 29 students (the smallest range of student group sizes currently included)—and among groups of 30 or more students. In unstabilized distributions, variance was higher for groups of 10 to 19 students and, to a lesser extent, for groups of 20 to 29 students than for the overall distribution (groups of 30 or more students) (figure 6). After stabilization, variance was lower for all student group sizes, and variance for small student groups was more comparable to that for larger student groups.

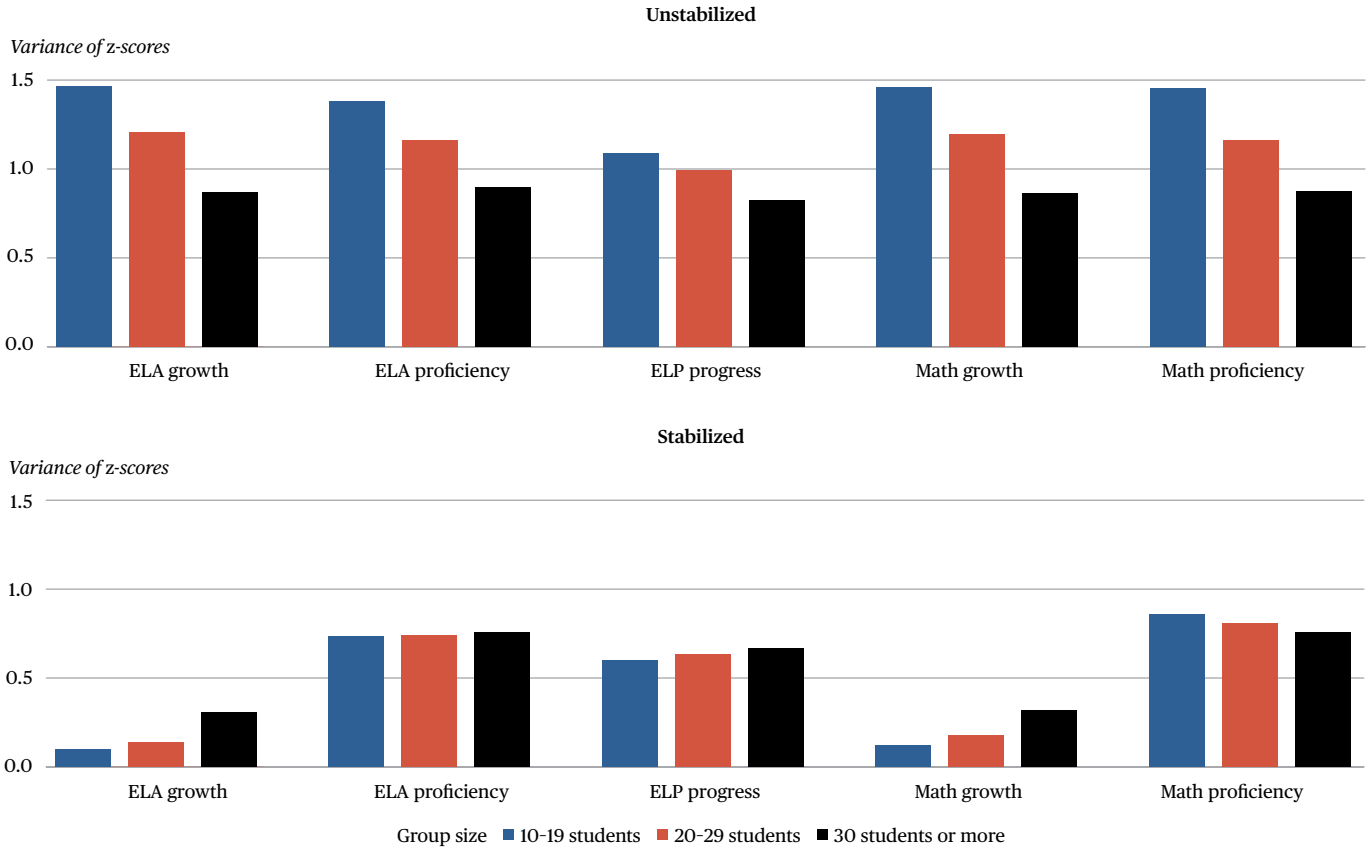
Second, stabilization should reduce overrepresentation of small student groups in the extremes of score distributions so that, on average, small student groups appear in the extremes of score distributions in approximately the same proportion as they appear across the entire score distribution. For example, if groups of 10 to 19 students account for 20 percent of student groups contributing to a specific subgroup-indicator score distribution, they should also account for about 20 percent of the student groups in the extremes of the score distribution. To explore this outcome, the study team analyzed the representation of groups of 10 to 19 students in the top and bottom 5th percentile of the score distributions before and after stabilization.

As expected, in unstabilized extremes of score distributions on all indicators, small student groups are overrepresented (figure 7). This is evident by comparing the unstabilized representation in the extremes (blue bars) with the expected representation based on the full performance distribution (black bars). Stabilization reduced overrepresentation of small student groups in the extremes for every indicator. This is illustrated by the red bars in figure 7, which show the representation of small student groups at the extremes of the distribution using stabilized scores and which are always smaller than the blue bars.

For the proficiency indicators, the representation of small student groups in the stabilized extremes was very close to their expected representation (the red bars are close in size to the black bars). Stabilization had a larger effect for the ELA and math growth indicators than for the proficiency indicators. This expected effect arises from differences in the construction of proficiency and growth indicators and in their resulting distributions and reflects the lower ratio of true performance to measurement error in growth indicators.[6] Student growth

6. An indicator constructed by taking the difference of two underlying measurements will often have a smaller "signal to noise" ratio than the underlying measurements. Here "signal" refers to the difference in true performance and "noise" refers to the difference in errors. This is because, for indicators where each score (measurement) must be greater than or equal to zero, the absolute value of the difference between the two scores will be smaller than either score if the second score is between half and double the first score. However, random error can be positive or negative, and the probability of being either is assumed to be equal. Therefore, the difference in errors will be greater than either error if the errors have different signs or if the second error is between half and double the first error. So, because change over time is often slow for academic indicators, "signal" will tend to decrease when taking the difference between two measurements and "noise" is less likely to do so.

Figure 6. Variance of z-score distributions for small groups of students is more aligned with the overall distribution after stabilization



ELA is English language arts. ELP is English language proficiency.

Note: Stabilized and unstabilized score data. Data were collected between school years 2015/16 and 2021/22 for proficiency rates, between 2016/17 and 2018/19 for growth, and between 2017/18 and 2021/22 for ELP progress. Data for school years 2019/20 and 2020/21 were omitted due to disruptions caused by the Covid-19 pandemic.

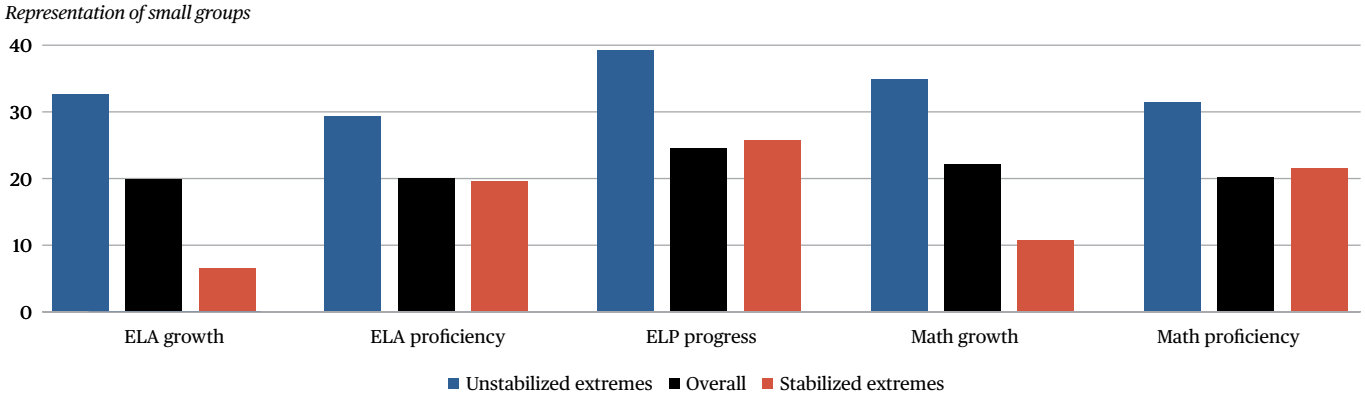
Source: Analysis of unstabilized score data aggregated and cleaned by the New Jersey Department of Education, also available online (New Jersey Department of Education, 2024).

percentiles have been shown in prior research to include a substantial amount of random error (McCaffrey et al., 2015; Wells & Sireci, 2020). This relatively high measurement error makes the indicator less statistically reliable and produces larger adjustments by the stabilization model.

For growth indicators, stabilization was substantial enough that small groups, especially those with 10 to 19 students, were underrepresented in the extremes of the stabilized score distributions. Underrepresentation of small student groups in the extremes of the stabilized distributions of growth indicators suggests that growth indicators are too noisy (too sensitive to measurement error) for groups of fewer than 20 students to confidently identify all such groups whose true performance is in the extremes. Stabilization produces results that are more accurate on average than unstabilized results. For noisy indicators like student growth percentiles, stabilization may lead to underrepresentation of small student groups in the extremes of the distribution. However, some underrepresentation will not make small student groups invisible to accountability processes as size cutoffs do. Additionally, decisionmakers can have greater confidence that student groups will be accurately designated as low performing using stabilization. Accuracy will improve in two ways: by increasing the correct identification of low-performing student groups and by avoiding incorrect identification of student groups that appear low-performing due to measurement error rather than true performance.

Figure 7. Stabilization better aligns the representation of small student groups of 10 to 19 students in the extremes of score distributions with their representation in the overall distribution

Representation in extremes of score distribution of groups of 10–19 students



ELA is English language arts. ELP is English language proficiency.

Note: Stabilized and unstabilized score data. Data were collected between school years 2015/16 and 2021/22 for proficiency rates, between 2016/17 and 2018/19 for growth, and between 2017/18 and 2021/22 for ELP progress. Data for school years 2019/20 and 2020/21 were omitted due to disruptions caused by the Covid-19 pandemic. The figure shows, for each indicator, the representation of small student groups of 10 to 19 students as a percentage of the unstabilized extremes (blue bar), overall (black bar), and stabilized extremes (red bar) of the distribution. In the case of the ELA growth indicator, for example, the figure shows that groups of 10 to 19 students make up over 30 percent of the extremes of the unstabilized distribution but only about 20 percent of the distribution overall. After stabilization, they make up less than 10 percent of the unstabilized extremes.

Source: Analysis of unstabilized score data aggregated and cleaned by the New Jersey Department of Education, also available online (New Jersey Department of Education, 2024).

Stabilization produced substantial changes in designations for Comprehensive Support and Improvement and Additional Targeted Support and Improvement

To understand how stabilization could affect decisionmaking processes and resource allocation, the study team applied NJDOE's process for designating schools for CSI and ATSI[7] to stabilized data for all test-based indicators. The study team compared CSI and ATSI designations made using composite scores constructed from unstabilized indicators—New Jersey's current approach—to designations made using composite scores constructed from stabilized test-based indicators. This illustrates the extent to which stabilization may affect CSI or ATSI designations if applied as recommended in this report.

The study team examined CSI designations based on a school's overall performance in the bottom 5th percentile of Title I schools in the state, as measured by a composite score. The team considered how frequently stabilization changed CSI or ATSI designations and for which student group sizes this was most likely to occur. Because stabilization reduces the overrepresentation of small student groups in the extremes of individual score distributions, it should reduce the percentage of schools for which small student groups led to designations of schools or student groups for improvement based on composite scores. Because New Jersey identifies schools and student groups for support and improvement based mostly on their relative performance, stabilization should produce a corresponding increase in the percentage of schools designated based on the performance of larger student groups. Finally, because a CSI designation is determined largely by schoolwide performance and an ATSI designation is determined by the performance of student groups in schools, the smaller student groups involved in an ATSI designation should see more changes relative to a CSI designation.

Smaller schools were less likely to be designated for Comprehensive Support and Improvement when test-based indicators were stabilized. For schools that met both the study's inclusion criteria for stabilization (see appendix B)

7. Described in NJDOE (2023).

and NJDOE's inclusion criteria for accountability, the study team did two things. First, it applied NJDOE's procedures for designating schools for CSI based on overall low performance. Second, it assigned CSI designations for the 2018/19 school year based on stabilized and unstabilized scores to compare how these designations were likely to change if NJDOE implemented stabilization. Of special interest were schools for which CSI designations differed when using unstabilized data and when using data that stabilized test-based indicators.

The procedure was applied to 2018/19 data because that was the most recent year of data that were not affected by the Covid-19 pandemic. NJDOE uses the following factors to designate schools for CSI:

1. The school's summative score—a weighted composite of scores from a set of contributing indicators, as defined in NJDOE's technical guide—is at or below the bottom 5th percentile of Title I schools.
2. The school has a four-year high school graduation rate at or below 67 percent.
3. The school is a Title I school and has been designated for ATSI for a low-performing student group for three or more consecutive years.

The study's simulated designations incorporated only the first criterion because stabilization was not applied to graduation rates, and New Jersey did not begin applying the third criterion until fall 2023.

Stabilization produced substantial changes in the list of schools designated for CSI. In total, 88 schools were designated for CSI using either unstabilized or stabilized test-based indicators (alongside unstabilized graduation rates and chronic absenteeism). Of these, 55 schools (63 percent) were designated using both unstabilized and stabilized indicators (table 2). Of the 72 schools designated for CSI using unstabilized indicators, 17 (24 percent) would not be designated if indicators were stabilized. Of the 1,908 schools not designated for CSI using unstabilized indicators, 16 (0.84 percent) would be designated if test-based indicators were stabilized.

To examine how CSI designations changed for schools of different sizes, the study team calculated the median school size for each type of CSI designation change (figure 8). The median school size overall was 475 students. The median size of schools that received a CSI designation using stabilized test-based indicators and that did not receive one using unstabilized test-based indicators was 551 students. The median size of schools that received a CSI designation using unstabilized test-based indicators and that did not receive one using stabilized test-based indicators was 479 students. That is, after stabilization, some larger schools replaced smaller schools among the schools designated for CSI.

These changes in CSI designation reflect changes in the distribution of composite scores after stabilization. As test-based indicator scores from small student groups are stabilized, on average, away from the extremes, the distribution shifts around the larger groups. When indicator scores are aggregated, the same shift occurs in the composite distribution, so that composite scores for smaller schools move away from the extremes of the

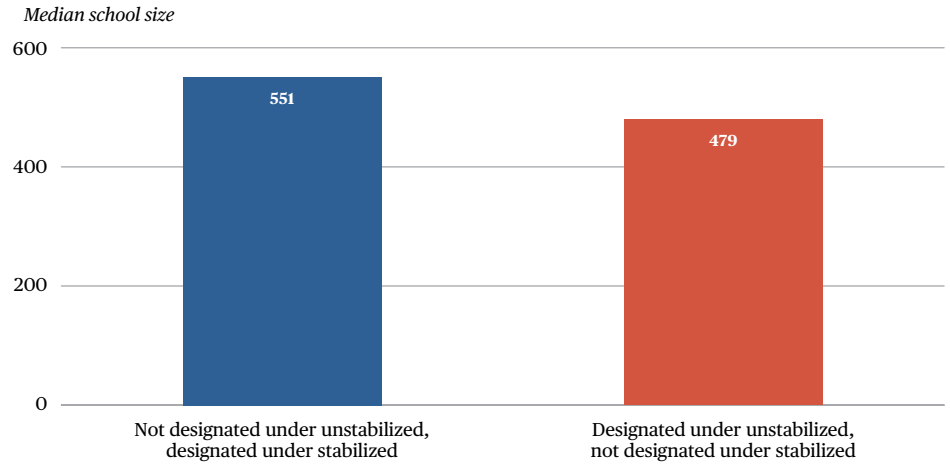
Table 2. Number of schools designated for Comprehensive Support and Improvement using stabilized or unstabilized test-based indicators (number of schools)

		Unstabilized test-based indicators	
		Not designated	Designated
Stabilized test-based indicators	Not designated	1,892	17
	Designated	16	55

Note: Stabilized and unstabilized score data. Data were collected between school years 2015/16 and 2021/22. Data for school years 2019/20 and 2020/21 were omitted due to disruptions caused by the Covid-19 pandemic.

Source: Analysis of unstabilized score data aggregated and cleaned by the New Jersey Department of Education, also available online (New Jersey Department of Education, 2024).

Figure 8. Median school size by Comprehensive Support and Improvement designation with stabilized and unstabilized data



Note: Stabilized and unstabilized score data. Data were collected between school years 2015/16 and 2021/22. Data for school years 2019/20 and 2020/21 were omitted due to disruptions caused by the Covid-19 pandemic.

Source: Analysis of unstabilized score data aggregated and cleaned by the New Jersey Department of Education, also available online (New Jersey Department of Education, 2024).

distribution while composite scores for larger schools are more stationary. The result is that more large schools occupy the extremes of the performance distribution not because their scores changed significantly under stabilization but because the distribution of indicator scores and, therefore, the distribution of composite scores changed around them. In effect, in the absence of stabilization, the outsized effect of measurement error in individual indicators on smaller student groups—and its aggregate effect on composite scores for schools—can effectively hide the low performance of some larger schools.

Additional Targeted Support and Improvement designation changes follow a pattern similar to that for Comprehensive Support and Improvement designation changes. The study also examined how ATSI designations change when stabilization is applied. Stabilization might matter more for ATSI designation than for CSI designation because an ATSI designation is based on the performance of groups of students, which are smaller than the whole school and therefore more sensitive to measurement error. ATSI designation rules vary substantially across states; New Jersey applies the ATSI designation to any school with a subgroup whose composite score is below the threshold that identifies schools for CSI. For consistency with the CSI analysis, the study team applied this analysis for the 2018/19 school year.

In total, 164 schools were designated for ATSI using either unstabilized or stabilized test-based indicators (alongside unstabilized graduation rates and chronic absenteeism). Of these, 95 schools (57 percent) were designated using both unstabilized and stabilized test-based indicators (table 3). Of 137 schools designated using unstabilized data, 42 (31 percent) would not be designated if indicators were stabilized, and they would be partly replaced by a smaller group of newly designated schools. Stabilization results in larger changes to the ATSI list than to the CSI list, as expected. The median size of the lowest performing subgroup in schools that received an ATSI designation was 294 students when stabilized test-based indicators were used and 238 students when unstabilized test-based indicators were used.[8] This difference is consistent with other results of this study,

8. In both cases, the median group size is quite large. This is partially because the stabilization process required multiple years of data, and CSI and ATSI processes required data from a minimum number of indicators. These requirements were more likely to be met by larger groups, resulting in a dataset with larger group sizes overall. Additionally, when reporting on group sizes for composite indicators, where contributing measures may have variation in group sizes, the largest group size is reported. This choice reflects an assumption that, in general, the smaller groups contributing to different indicators within the same subgroup and year are likely to be a subset of the largest group.

Table 3. Number of schools designated for Additional Targeted Support and Improvement using stabilized or unstabilized test-based indicators (number of schools)

		Unstabilized test-based indicators	
Not designated		Not designated	Designated
Stabilized test-based indicators	Not designated	1,816	42
	Designated	27	95

Note: Stabilized and unstabilized score data. Data were collected between school years 2015/16 and 2021/22. Data for school years 2019/20 and 2020/21 were omitted due to disruptions caused by the Covid-19 pandemic.

Source: Analysis of unstabilized score data aggregated and cleaned by the New Jersey Department of Education, also available online (New Jersey Department of Education, 2024).

reflecting that stabilization reduces the number of small student groups in the extremes of score distributions by reining in outliers that are driven by random error.

Patterns in Comprehensive Support and Improvement and Additional Targeted Support and Improvement designation changes are aligned overall. Overall, stabilization changed CSI and ATSI designations in a similar way. Generally, but not uniformly, stabilization moved small groups of students out of the extremes of score distributions for individual indicators and thereby changed the distribution of composite indicators, resulting in fewer small schools and small groups of students designated for support and improvement. Correspondingly, larger student groups were more likely to be designated for support and improvement after stabilization, not because stabilization had a large effect on their individual or composite indicator scores but because scores from smaller groups of students shifted around them. This had the effect of improving detection of low-performing large schools and student groups by stabilizing the performance of smaller schools and smaller student groups, for which measurement error has a larger effect on unstabilized scores.

Implications

This report shows that a simple stabilization model that assumes normally distributed errors is well-suited for improving the reliability of New Jersey's test-based performance indicators, which are well-aligned to model assumptions. For indicators that do not align with the assumptions of this model, implementing different modeling choices of varying complexity may improve suitability. For example, for indicators with highly skewed distributions, modelers may transform the data to a more normal distribution prior to stabilization.[9] Modelers may also adjust the model's assumptions about the distribution of scores to more closely align with the data by allowing for skew or for more extreme scores.[10] Additionally, results may be improved using more complex models, such as those that incorporate individual-level data, those that stabilize across multiple indicators within a subgroup, or those that are specifically designed to address bounded distributions.[11]

States wishing to apply stabilization may consider a variety of modeling options to meet their needs. This might involve applying a simple stabilization model only to test-based indicators or applying slightly different models to different indicators. Applying a simple stabilization model to test-based indicators can improve the accuracy of both the individual indicators and the composite indicators that states use in accountability designations.

9. This may be accomplished using a Box-Cox transformation. These transformations are supported by a variety of statistical software, making them easy to implement. They are also reversible, so the adjusted scores can be easily returned to their original scale.
10. This may be accomplished by selecting different prior distributions, such as gamma distributions (which allow for skew) or students' *t* distributions, which allow for more extreme scores.
11. However, due to their complexity and the large amount of data required to produce reliable results, these may not be feasible for state education agencies to implement.

Stabilization mitigates two types of error in accountability designations. The first is erroneously designating student groups as in need of support (a false positive); the second is failing to designate a student group as in need of support (a false negative). In both cases, resources and support are not allocated as intended. Stabilization reduces measurement error in scores based on student group size, overall score distributions, and historical performance for a group so that states can be more confident in their accountability designations. By reducing measurement error, stabilization substantially improved the reliability of scores from groups of 10 to 19 students, suggesting that NJDOE could use it as a tool to lower their size threshold for including student groups in accountability designations, thereby increasing the equity of their accountability system.

Overall, based on theory, prior evidence, and the findings of this study, it is reasonable to conclude that stabilization can improve both accuracy and equity in test-based indicators used for accountability and diagnostic purposes. States may consider applying stabilization as a tool to lower the size threshold for the inclusion of student groups in accountability processes without conern of overrepresentation of those groups in accountability designations.

References

Forrow, L., Starling, J., & 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

McCaffrey, D. F., Castellano, K. E., & Lockwood, J. R. (2015). The impact of measurement error on the accuracy of individual and aggregate SGP. *Educational Measurement: Issues and Practice*, *34*, 15-21. https://doi.org/10.1111/emip.12062

New Jersey Department of Education. (2023). *2021-2022 technical guide to Every Student Succeeds Act (ESSA) summative ratings and the identification of schools in need of support and improvement*. https://www.nj.gov/education/title1/accountability/docs/22/2021-22_ESSA_Technical_Guide_SummativeRatings_Identification.pdf

New Jersey Department of Education. (2024). *Accountability.* Title I, Part A. https://www.nj.gov/education/title1/accountability/

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954-55, Vol. I*, 197-206.

Wells, C.S., & Sireci, S.G. (2020). Evaluating random and systematic error in student growth percentiles. *Applied Measurement in Education, 33*(4), 349-361. https://doi.org/10.1080/08957347.2020.1789139