Institute of Education Sciences

Regional Educational Laboratory Mid-Atlantic

At Mathematica

Why School Accountability Systems Disproportionality Identify Middle Schools' SWD Subgroups for TSI

REL 2021-049
U.S. DEPARTMENT OF EDUCATION

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

Why school accountability systems disproportionately identify middle schools' SWD subgroups for TSI

Lauren Forrow and Kevin Kelly

October 2020

Key findings

Middle schools are identified for targeted support and improvement (TSI) at a high rate in two mid-Atlantic states based largely on the performance of their students with disabilities (SWDs) subgroups. We examined publicly available data from these states to understand the reasons for this high rate. Specifically, we investigated whether state-set minimum n-size requirements for an indicator to be included in the accountability system made it more likely for middle schools' SWD subgroups to be identified for TSI, or whether middle schools' SWD subgroups are genuinely lower-performing than SWD subgroups in elementary or high schools. Our analyses suggest a combination of two reasons for the higher rate of TSI identification in middle school SWD subgroups:

- SWD subgroups in middle schools are more likely than those in elementary schools and high schools to meet minimum sample size requirements for academic proficiency indicators, which are commonly based on statewide assessments.
- At all school levels, SWD subgroups perform poorly relative to the overall student population on academic proficiency indicators.

Why this study?

The Every Student Succeeds Act (ESSA) requires states to identify schools in need of Comprehensive Support and Improvement (CSI), Targeted Support and Improvement (TSI), and Additional Targeted Support and Improvement (A-TSI) based on the performance of their students.[1] CSI schools are those with the lowest performance across all students in the state; TSI and A-TSI schools are those in which at least one subgroup of students, such as students from families with low income or English learner students, is low performing.

Some states in the Regional Educational Laboratory (REL) Mid-Atlantic region have noted that, compared to other school levels, middle schools are more often identified for TSI, A-TSI, or both, often based on the performance of their students with disabilities (SWDs) subgroup. For example, in one state, six out of nine schools identified for TSI (approximately 67 percent) were middle schools identified for their SWD subgroup. This pattern was prominent enough in two REL Mid-Atlantic states that State Education Agency representatives from these states asked the REL to investigate. Specifically, the states wanted to know whether these results reflect systematically low performance among middle school SWDs or whether the frequent identification of these subgroups as requiring support is an unintended consequence of the rules underlying states' school accountability systems. The latter possibility is bolstered by research suggesting that the rules that make up these systems affect the sets of schools and subgroups identified for support. For example, reducing the minimum number of students required to hold a school accountable, or a subgroup within a school accountable, leads to more schools being included in the accountability system and therefore eligible to be identified for CSI, TSI, or A-TSI (Cardichon, 2016). As another example, requiring A-TSI schools to first be identified as TSI, an approach used in at least nine states and Puerto

[1] CSI schools are the lowest-performing 5 percent of Title I schools in the state, high schools with a four-year adjusted cohort graduation rate at or below 67 percent, and Title I schools previously identified as A-TSI that did not improve. TSI schools are those with one or more consistently underperforming subgroup, as defined by the state. A-TSI schools are those in which one or more subgroups of students performs at or below the performance of all students among CSI schools.

Rico, can result in some schools that would otherwise qualify for A-TSI not receiving this support (Dworkin & Hyslop, 2019).

As states look to future rounds of accountability designations under the Every Student Succeeds Act, they aim to continuously improve the ways in which schools are identified for support by thoughtfully examining the rules underlying their accountability systems. To that end, this study examined how one rule—the minimum n-size requirement, which specifies how many students must have scores for an indicator to be included in the school's accountability score—may affect the group of schools identified for TSI in two states with overrepresentation of middle schools among their TSI schools.[2] This research also provides insight to other state education agency (SEA) leaders beyond those from the states included in the study. All state school accountability systems include a minimum n-size requirement, so understanding any implications of such a requirement is beneficial to state leaders aiming to refine their accountability systems. Further, researchers who conducted semi-structured interviews with SEA officials across seven states found that most officials were surprised by the number of TSI/A-TSI schools identified in their states and that SWD subgroups were the most common subgroup leading schools to be identified for these supports (Rentner et al., 2019). Although this research does not point to common nationwide TSI identification for SWD subgroups at middle schools specifically, it suggests that understanding the role the minimum n-size requirement plays in identifying SWD subgroups for TSI or A-TSI will inform SEA officials across the nation as they aim to improve the ways in which their states' accountability systems identify schools and subgroups in need of support.

School accountability systems

States identify TSI schools through accountability systems that typically summarize school performance in a single score. Accountability rules combine information from several indicators that describe different dimensions of students' performance. These dimensions may include academic proficiency, academic growth, graduation rates, English learner students' progress toward English fluency, and other measures of school quality and students' success. The state calculates a score for each subgroup in every school based on the indicators in the state's school accountability system; in the remainder of the report, we refer to these metrics as accountability indicators (box 1). If the number of students in a school with indicator data from a particular subgroup is less than the state's minimum n-size requirement, that subgroup does not receive a score for that indicator. For example, consider a school with math proficiency data for 15 Black or African American students. If the state's minimum n-size is 10 students, the Black or African American student subgroup will receive an indicator score for math proficiency. If the state's minimum n-size is 20 students, however, the Black or African American student subgroup will not receive an indicator score for math proficiency.

Box 1. Accountability indicators used in this study

Accountability indicators are metrics on which states assess schools to make accountability designations. States use several accountability indicators across various dimensions. In this study, we examined six accountability indicators, which we categorize into four sub-categories:

1. **Proficiency indicators** measure the percentage of students who scored "proficient" on standardized exams. There are two proficiency indicators:
    - **Math proficiency**, based on standardized math exams.
    - **English language arts (ELA) proficiency**, based on standardized ELA exams.

[2] For simplicity, here and through the remainder of this report, we refer to both TSI and A-TSI schools as TSI schools.

2. **Academic growth indicators** measure the average change in students' performance on standardized exams from the previous year. Academic growth indicators are typically measured in standard deviations. There are two growth indicators:
   - **Math growth**, based on the change in performance on math exams.
   - **ELA growth**, based on the change in performance on ELA exams.
3. The four-year adjusted cohort **graduation rate indicator**, hereafter called the graduation rate indicator, measures the percentage of students who graduated from a high school within four years of enrolling in grade 9 at the school.
4. The **chronic absenteeism indicator** measures the percentage of students who were chronically absent, defined by both states included in this study as missing 10 percent or more of the days in which they were enrolled in the school. However, to align the interpretation of the indicator with the interpretations of the remaining accountability indicators, where higher values connote positive outcomes, states commonly transform this indicator when performing accountability calculations. We follow the states in this practice, presenting results for the percentage of students who were not chronically absent—that is, they attended school at least 90 percent of the time they were enrolled. We refer to this as the "regular attendance rate."

Throughout the analyses, we rely on individual states' calculations of these indicators as provided in public use data sets. This decision is appropriate for this study because, although different states may define each measure differently, we use each state's preferred definition in the analyses for that state. Because this study focuses on how accountability indicators are used in schools' accountability systems and does not consider how alternative definitions of these indicators might affect accountability, we do not provide further details regarding the specific calculation of each indicator in each state.

Through a series of rules they submit for approval by the U.S. Department of Education, states evaluate the overall performance of each subgroup in a school by combining indicator scores to create a composite score for that subgroup. Importantly, even when the state's rules mandate the inclusion of particular indicators, for an individual school and subgroup this composite score includes only indicators for which the subgroup meets the minimum n-size requirement. In many schools, some student subgroups meet the minimum n-size requirement for some but not all indicators, so their overall performance is evaluated on only a subset of indicators.[3] For example, an elementary school might have enough SWDs across grades K to 5 to meet the minimum n-size requirement for an indicator such as regular attendance, but not enough SWDs in tested grades (grades 3 to 5) to meet the minimum n-size requirement for proficiency in English language arts (ELA) or math. Schools are ultimately identified for TSI based on the overall performance of each of their student subgroups. If at least one subgroup in a school is deemed low performing, as defined by each state, the school is identified for TSI.

This study expands the literature on the tradeoffs and implications of setting a minimum n-size requirement to examine the real-world implications of the minimum n-size requirement on identifying TSI schools in two states. Some scholars argue that states' accountability systems can only support schools with struggling student groups if those student groups are included in the accountability system, advocating for as small an n-size as possible (Cardichon, 2016). In addition, an analysis of historical data in California showed that reducing the minimum n-size requirement from 100 to 20 led to six times as many schools reporting results for African American students in the state's accountability system (Hough et al., 2016). From this point of view, low minimum n-sizes expand states' understanding of the performance of student subgroups. However, lowering the minimum n-size requirement may also reduce the reliability and validity of reported results or compromise students' privacy (Seastrom, 2017; Data Quality Campaign, 2017; Sabia & Cortiella, 2017).

[3] Some states require that subgroups have indicator scores for a certain number of indicators in order to evaluate overall performance, but many states do not include this requirement.

Because of the sensitivity surrounding school accountability determinations, the two states whose TSI identification results prompted this research requested anonymity in this public-facing report. Thus, we refer to these states as State A and State B. More context about these states, including information about their TSI identification results and their accountability systems, is provided in box 2.

Box 2. States included in this study

In **State A**, two-thirds of the schools identified for TSI were middle schools identified for their SWD subgroup. The minimum n-size requirement in State A is 15, and all subgroups in a school are eligible to be identified for TSI as long as the subgroup has an indicator score for at least one indicator. That is, 15 or more students from the subgroup must have indicator data for at least one indicator for the subgroup to be included in the state's school accountability system. In State A, a subgroup's overall accountability score is a weighted average of the raw indicator scores, and the weight associated with each indicator varies by school level (appendix B). Weights are adjusted based on the number of indicators for which the subgroup has a score.

State B also identified a larger number than expected of middle schools for TSI because of their SWD subgroups; roughly 10 percent of all schools identified for TSI. State B's minimum n-size requirement is 20, and subgroups in a school are eligible to be identified for TSI only if they have indicator scores for at least three of four core indicators. State B's accountability calculation, which is more complex than State A's calculation, follows these three steps:

1. It categorizes schools based on the number of indicators for which they meet the minimum n-size requirement.
2. It transforms raw indicator scores to a common scale.
3. It calculates the overall accountability score as a weighted average of the transformed indicator scores.

Appendix B describes the calculation in full.

Research questions

This study explored whether school accountability rules, particularly minimum n-size requirements, are associated with the disproportionate identification of middle schools' SWD subgroups for TSI. Two questions guided this research:

1. Are SWD subgroups at middle schools more likely to meet minimum n-size requirements for academic proficiency indicators than SWD subgroups at other school levels?

Because middle schools tend to be larger than elementary schools, their SWD subgroup might be more likely to meet minimum n-size requirements across all indicators than SWD subgroups in elementary schools. In addition, more middle school grades take statewide assessments than high school grades, so a middle school's SWD subgroup might be more likely to meet minimum n-size requirements for the ELA and math proficiency indicators than SWD subgroups in high schools. Although ELA and math growth indicators are also based on statewide assessments, we mainly focus on academic proficiency indicators because they are included in accountability scores for schools at all levels, whereas academic growth indicators are included in accountability scores only for elementary and middle schools. Other indicators that are not linked to state assessments, like regular attendance, tend to have more robust sample sizes and are thus less affected by minimum n-size requirements.

2. How does the performance of SWD subgroups differ from the performance of the all students group on each accountability indicator? Is the difference larger in middle schools than in elementary or high schools?

TSI schools are those in which one or more student subgroups performs poorly relative to the average performance across the state across accountability indicators. To investigate why, in some states, middle schools' SWD subgroups are disproportionately identified for TSI, we must compare SWDs' average performance on each accountability indicator with the average performance on these indicators across all students. We perform this comparison by school level to gauge whether the difference in performance between SWDs and all students is larger for middle schools than other school levels; if so, performance on accountability indicators might partly explain why more middle schools are identified for TSI.

Box 3 and appendix A describe the data and methods used to explore these research questions.

Box 3. Data sources and methods

Data sources. The data set of accountability information was compiled from public data portals. This data set included the score and number of students included in the score for each of six accountability indicators for which 2017/2018 school year data were available from both states: ELA proficiency, math proficiency, ELA growth, math growth, chronic absenteeism, and four-year graduation rates (for high schools only).

Sample. The public data files include information for all public schools with accountability scores from the 2017/2018 school year. In the public data files, scores are missing when the sample did not meet the minimum n-size in that state.

Methodology. Each school is classified as elementary, middle, or high based on the grade levels represented in the school. For consistency, the same rules are used in both states. We considered a school to be an elementary school if it had students in grades K to 5 and no students above grade 6; a middle school if it had students in grades 7 and 8 but not below grade 5 or above grade 9; and a high school if it had students in grades 10 to 12 and no students below grade 8. For schools that included grades in more than one of the three levels, such as K to 8 schools, we classified them based on the majority of grades represented.

To answer research question 1, we calculated the number and proportion of schools meeting minimum n-size requirements for ELA and math proficiency indicators for the SWD subgroup for each combination of state and school level. For example, we calculated the number of elementary schools in State A where the SWD subgroup met the minimum n-size requirement for each proficiency indicator. We determined, for each school, whether at least one academic proficiency indicator met the state's minimum n-size requirement. We then calculated, for each combination of state and school level, the percentage of schools in which the SWD subgroup's accountability score included at least one academic proficiency indicator.

To answer research question 2, we calculated the mean of school-level scores for each indicator separately by state, school level, and subgroup (all students versus SWDs), weighting by the number of students at each school. We then took the difference between the average score in the all students group and the average score in the SWDs group to obtain an estimate of the performance gap for each indicator, state, and school level.

For both research questions, we examined differences across and within school levels. Because our analysis examined school-level data rather than student-level data, we were unable to calculate accurate measures of statistical uncertainty for these differences. Statistical significance testing, which relies on such measures of uncertainty, is therefore not feasible.[4] Instead, as a heuristic, we note **substantial differences**, which are differences of at least 20 percentage points.

More detailed information on the methodologies is available in appendices A and B.

Limitations. This study relied on publicly available data, which introduced several limitations. First, publicly available data are at the school level, and student-level data are essential to more comprehensively examine the relationships among minimum n-size requirements, other features of accountability rules, and TSI identification. As noted above, student-level data are also required to calculate statistical uncertainty accurately, so without these data we are not able to conduct statistical significance tests of differences across subgroups or school levels.

[4] Although it would be possible to conduct statistical significance tests at the school level—for example, tests of the difference in proportions of schools meeting minimum n-size requirements for academic proficiency measures among elementary, middle, and high schools—we chose not to do so because, in our view, such tests do not address the research question of interest and because the statistical literature increasingly urges caution in relying on significance tests (for example, see Wasserstein & Lazar, 2016).

Publicly available data sets also suppress information based on small sample sizes and other privacy concerns, which further limited the data available for analysis. The public use files provided adequate information to assess whether schools' indicator scores were suppressed because the underlying sample size did not meet the state's minimum n-size requirement or because the average score was so low that it would violate students' privacy to report it. Scores that are suppressed due to low sample size would not be included in accountability calculations, so their absence from the data file does not bias our results. However, scores that are suppressed due to low average score would be included in accountability calculations, and they account for a moderate proportion of the suppressed ELA and math proficiency scores for SWD subgroups in State B. After consulting with State B, we conducted a sensitivity test where we imputed indicator scores as 0 if the indicator score was suppressed but the corresponding sample size met the state's minimum n-size requirement. The inclusion of these schools' subgroups' scores as 0 does not affect our conclusion (figure C.2 in appendix C).

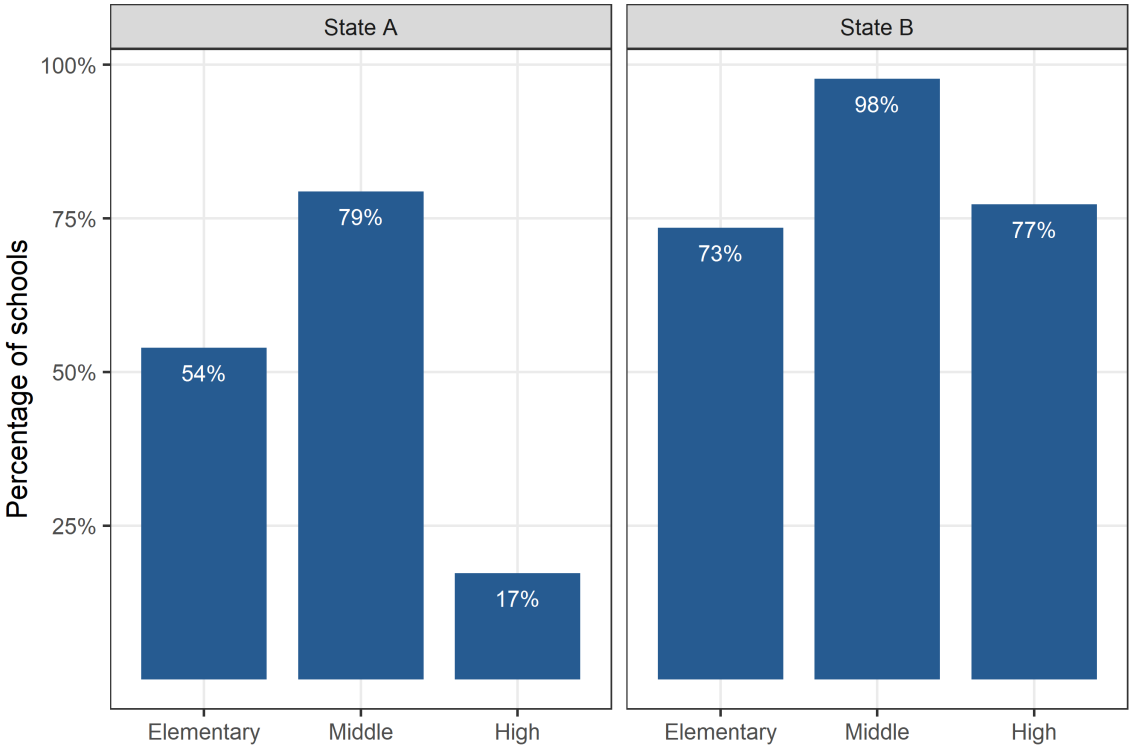
Second, because adequate data were available for only six accountability indicators, we could not comprehensively examine all dimensions of the performance of middle schools' SWD subgroups. A more complete investigation would have incorporated these sources of data. Finally, because states have only just begun to implement their ESSA plans, only one year of TSI identification data was available.

Findings

Students with disabilities subgroups at middle schools are more likely to meet minimum n-size requirements for academic proficiency indicators than are students with disabilities subgroups at other school levels

We compared the frequency with which schools' SWD subgroups meet minimum n-size requirements across school levels and states. The results show that middle schools' SWD subgroups are substantially more likely to meet the minimum n-size requirement for at least one academic proficiency indicator than SWD subgroups in elementary or high schools (figure 1).

Figure 1. Students with disabilities subgroups at middle schools are more likely to meet the minimum n-size requirement for academic proficiency indicators than these subgroups are at other school levels



Note: The heights of the bars give the percentage of schools meeting the minimum n-size threshold in that state based on the maximum number of students with scores for English language arts and math proficiency indicators.
Source: Publicly available data from the departments of education in States A and B.

In State A, almost 80 percent of middle schools meet the minimum n-size requirement for at least one academic proficiency indicator, compared with 54 percent of elementary schools and 17 percent of high schools. Although a higher proportion of schools at all levels meet the minimum n-size requirement for at least one academic proficiency indicator in State B, middle schools are more than 20 percentage points more likely to meet this threshold than elementary or high schools. These results indicate that overall accountability scores for middle school SWD subgroups are substantially more likely to incorporate academic proficiency information than are overall accountability scores for elementary or high schools.

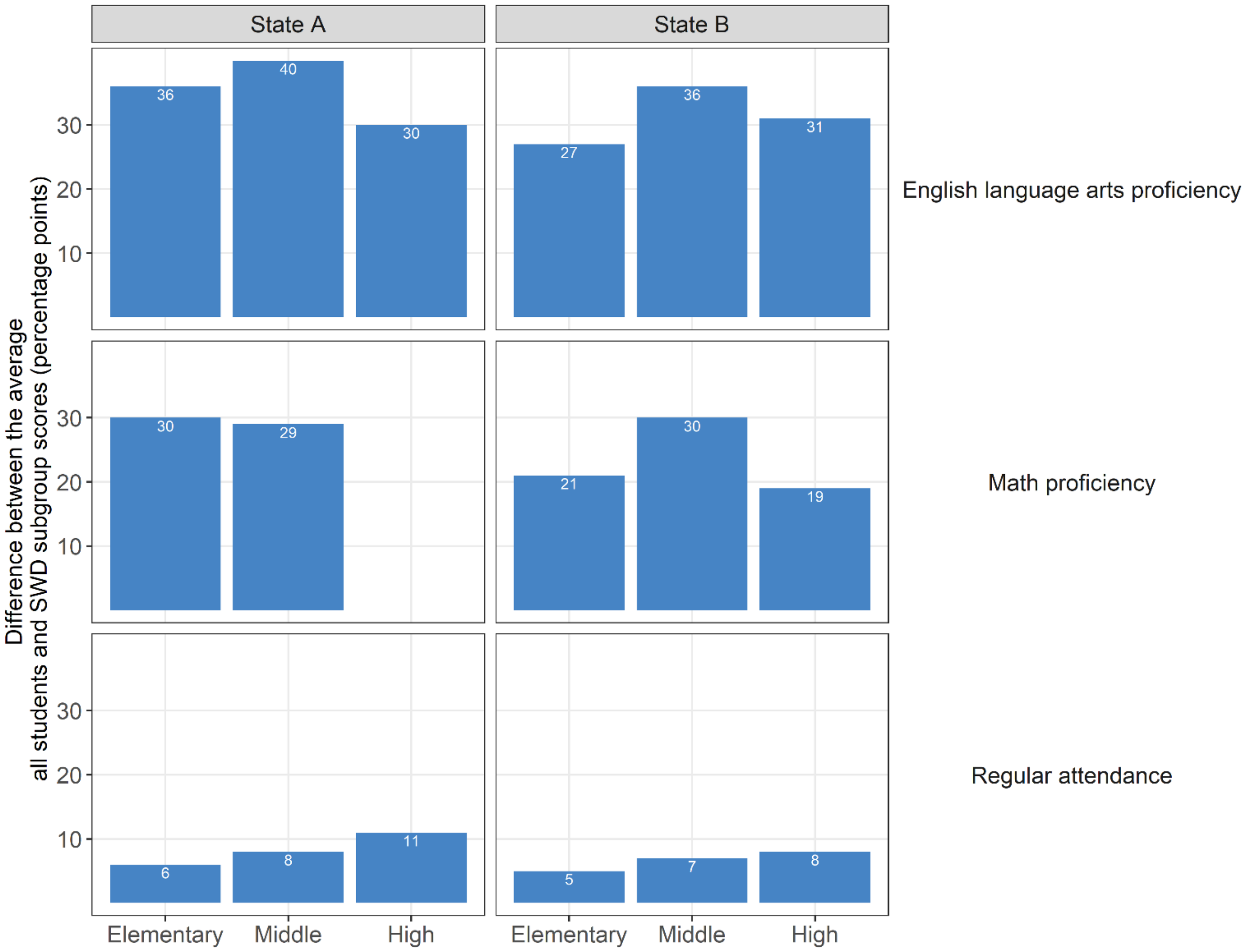
Students with disabilities subgroups perform poorly on academic proficiency indicators relative to all students

To determine whether poor performance can explain disproportionately high TSI identification rates, we must compare SWD subgroups with the all students group at each school level. We must also gauge whether the performance gap between the all students group and SWD subgroups is greater at some school levels than others. If middle school SWD subgroups perform worse relative to the all students group than SWD subgroups do at other school levels, this performance difference could also contribute to the high TSI identification rate for middle schools' SWD subgroups.

Because the accountability score is a weighted average of many indicators, it is equally important to perform this comparison for each indicator, not just the academic proficiency indicators. If middle school SWD subgroups perform worse than SWD subgroups at other school levels, relative to the all students group, on indicators other than academic proficiency, the higher probability of meeting the minimum n-size for academic proficiency indicators would only partly explain the high rates of TSI identification for middle schools' SWD subgroups. Lower performance on other indicators would be another part of the explanation.

The percentage point difference between the average all students group score and SWD subgroup score for each school level is calculated for three indicators: ELA proficiency, math proficiency, and regular attendance (figure 2). Results for the remaining indicators are included in figure C.1 in appendix C. All the bars in the figure are positive because the all students group consistently outperformed the SWD subgroup across school levels and indicators. Differences are particularly large, however, on the ELA and math proficiency indicators. For example, in State A, 57 percent of all students in elementary schools were proficient in ELA, compared to 20 percent of SWDs at this school level, leaving a difference (with rounding) of 36 percentage points. (See table C.1 in appendix C for average academic indicator scores by state, school level, and subgroup.) Differences are smaller—between 5 and 11 percentage points—for regular attendance.

Figure 2. Performance of students with disabilities relative to all students does not vary widely across school levels



Note: The height of each bar is calculated as the average indicator score for all students minus the average indicator score for the SWD subgroup across schools meeting the minimum n-size criterion based on all indicators. The average math proficiency score for the State A high school SWD subgroup was redacted in the original data, so no value is available for this score.

Source: Publicly available data from the departments of education of States A and B.

The other indicators that we examined (ELA growth, math growth, and graduation rate) follow a similar pattern to regular attendance: SWD subgroups consistently perform worse than the all students group but not to the same extent as for the proficiency indicators, and not substantially (figure C.1 in appendix C). In both states, SWD subgroup performance on the academic growth indicators is within 10 percentage points of the all students score. Although the differences are slightly larger for the graduation rate indicator, they are smaller than the differences for the proficiency indicators in high schools, the only school level for which graduation rates are included in accountability scores. For example, in high schools in State A, the difference in graduation rates between the SWD subgroup and the all students group is 17 percentage points, compared to a 30 percentage point difference in ELA proficiency scores.

Students with disabilities subgroups in middle schools appear to perform worse on academic proficiency indicators, relative to the all students group, than students with disabilities at other school levels

At first glance, differences in performance appear to be somewhat larger for middle schools, but this pattern is not completely consistent (see figure 2). The gap is larger for middle schools on ELA proficiency in both states, although not substantially. However, for math proficiency and regular attendance, the performance gap could be larger, smaller, or the same for middle schools compared to other school levels, with no clear pattern across states. The remaining indicators (appendix C) reveal similarly inconsistent patterns across school levels and states. Importantly, without student-level data we cannot accurately calculate the statistical uncertainty in indicator scores, which would also affect our understanding of how the performance gap varies by school level. Without this information, we are not confident in our ability to determine whether performance gaps as similar as 40 and 36 percentage points, in ELA proficiency in State A for middle and high school students, respectively, reflect true differences or random error. In addition, as noted above, the performance gap in middle schools is never substantially larger—that is, it is never larger by 20 percentage points or more—than the gap in elementary or high schools. This difference in average indicator scores is likely to contribute to the TSI identification results among SWD subgroups in middle schools, but, because of the uncertainty in the estimates, it is not clear how much.

In contrast to differences in the performance gap, differences across school levels in the percentage of schools meeting the minimum n-size requirements for academic proficiency indicators are much larger and more concrete. In State A, for example, there is a 25 percentage point difference in the percentage of elementary and middle schools meeting the minimum n-size requirement for academic proficiency indicators; however, there is at most a 10 percentage point difference in the performance gap on academic proficiency indicators at these school levels. The difference in the percentage of schools meeting the minimum n-size requirement is both larger (meeting our threshold for a substantial difference) and less subject to statistical error because it is not based on state assessments. Differences in the percentage of schools meeting the minimum n-size requirement for academic proficiency indicators are therefore more likely to drive the identification of middle schools’ SWD subgroups for TSI.

Other features of accountability systems could moderate the role of minimum n-size requirements

The main results of this study indicate that differences in the rates at which SWD subgroups meet minimum n-size requirements for academic proficiency measures could increase their probability of TSI identification, relative to SWD subgroups at other school levels. The accountability rules in States A and B suggest that other features of the accountability system can reduce the impact of this phenomenon. As we noted previously, State B excludes from its accountability system the schools that do not meet the minimum n-size requirement for a minimum number of specified indicators, including academic proficiency indicators. In State B, the differences across school levels in the percentage of schools meeting the minimum n-size requirement for at least one academic proficiency indicator are smaller than in State A (see figure 1), and the proportion of middle schools identified for TSI for their SWD subgroups is also lower (10 percent of TSI schools compared to 67 percent in State A). State B’s exclusion criteria could facilitate greater consistency in the proportion of schools meeting the minimum n-size requirement for at least one academic proficiency indicator across school levels.

State B also groups schools based on the number of accountability indicators for which they meet the minimum n-size requirement and compares performance within rather than across groups. Categorizing schools in this way could further mitigate the consequences of different rates of meeting minimum n-size requirements: a school that meets the minimum n-size for both academic proficiency indicators for its SWD subgroup, and therefore might have a lower overall accountability score, is not compared to a school that did not meet the minimum n-size for either academic proficiency indicator for its SWD subgroup.

Although not as directly linked to TSI identification results, transforming raw indicator scores onto a common scale could also improve the statistical reliability of State A's accountability system.

Implications

This study provides evidence suggesting that SWD subgroups in middle schools may be disproportionately identified for TSI in two REL Mid-Atlantic states, in part because they are more likely to meet minimum n-size requirements for academic proficiency indicators, on which SWD subgroups perform poorly relative to the all students group. The first question that this finding prompts is whether universal minimum n-size requirements are too rigid. In a small school, a small absolute number of students can account for a large proportion of the student population, and excluding their scores from accountability calculations can represent a substantial loss of information. However, a subgroup that represents a moderate proportion of a small student population could still be so small that it is statistically unreliable; average scores for groups of three or four students, even if they account for more than 10 percent of enrollment, could vary wildly from one year to the next. With no clear alternative that accounts for both statistical reliability and size relative to total enrollment, strict minimum n-size requirements are the most straightforward way to ensure statistically reliable results. However, other features of school accountability systems could moderate the influence of minimum n-size requirements, so we focus on these features as avenues SEA officials can consider when addressing similar accountability concerns.

One approach that SEA officials could consider implementing in their states' accountability systems is stratification, whereby states categorize schools into groups based on the number of proficiency indicators for which the school met the minimum n-size requirement. Stratification ensures that, when identifying schools for TSI, schools with accountability scores that incorporate academic proficiency indicators are not compared with schools with scores that do not incorporate these indicators. Because the evidence suggests that, at least for the SWD subgroup, including academic proficiency indicators is likely to lower accountability scores, comparing schools with like numbers of proficiency indicators will avoid some of the inherent issues with dropping groups that do not meet the minimum-n-size requirement. In the future, researchers could work collaboratively with SEA officials to explore the effects of different implementations of stratification on the number of proficiency indicators to further inform potential refinements to states' accountability systems.

Although a stratification approach could help address perceived unfairness in comparing schools with accountability scores that incorporate different indicators, it would not resolve a larger, perhaps more significant question that this study's results suggest. In essence, this study shows that middle schools are more likely to be identified for TSI largely because more information is available about their performance; that is, middle schools are more likely than elementary and high schools to include information on academic proficiency. This finding could be considered positive—SEA officials can know that accountability systems are identifying the schools that are most certain to need support. However, they could equally be concerned that elementary and high schools requiring support do not receive it if, because of small sample sizes, their accountability indicator scores are too unreliable to use as the basis of TSI determinations.

From this perspective, the study's findings suggest that SEA officials and researchers should continue to collaboratively explore methods and strategies that accurately assess performance based on small numbers of students. Findings from these explorations could further influence how SEA officials set the rules underlying their states' accountability system. This work is particularly relevant in the current context, as states must reconsider accountability rules to acknowledge the dramatic disruption of the 2019/2020 school year due to COVID-19. As they consider revisions, SEA officials could consider the options that address assessing performance based on small numbers of students.

One possible path of exploration is to use an empirical Bayesian approach to increase the precision of small subgroups' accountability scores, allowing states to assess all subgroups on all indicators without unduly increasing the risk of distorting accountability results with statistical noise. Unlike the standard approach in school accountability calculations, where each school's or subgroup's score is based only on data from that school or subgroup, empirical Bayesian methods introduce structured assumptions that pool information to improve the precision of estimates for schools or subgroups with comparatively little data. Such methods are already common in teacher value-added models (see, for example, Walsh et al., 2014). However, state education agencies might not be prepared to implement this method in a school accountability setting, or they might be concerned that this approach would be confusing to much of the general public, limiting the transparency of their accountability systems.

Another possible solution is to aggregate student data in an effort to include more subgroups that do not meet the minimum n-size requirement in the accountability system. For example, combining data across grade levels, over years, or across underperforming subgroups may increase the representation of subgroups in states' accountability systems (Gordon, 2017). Each of these approaches has implications for how the state would support schools identified as underperforming. For example, suppose a state aggregates data across three years for each school and subgroup to achieve adequate sample size in subgroups. In this system, accountability determinations in 2020 are based on data from the 2018, 2019, and 2020 school years, even though different students attend the school in each year. Under such a system, should an elementary school identified as needing additional support for its SWD subgroup in 2020 receive that support, even though most of the students included in the aggregated accountability data—in a K–5 school, those in third grade or higher in 2018—now no longer attend the school?

The question of how to assess performance accurately for subgroups with small numbers of students, and the range of possible solutions, calls for thorough and careful methodological investigation by researchers working collaboratively with SEA officials who offer key contextual knowledge of the many facets of states' school accountability systems to ensure that all students receive the necessary support.

References

Cardichon, J. (2016). *Ensuring equity in ESSA: The role of n-size in subgroup accountability*. Alliance for Excellent Education. https://all4ed.org/wp-content/uploads/2016/06/NSize.pdf

Data Quality Campaign (2017). *Understanding minimum n-size and student data privacy: A guide for advocates*. https://dataqualitycampaign.org/wp-content/uploads/2017/06/DQC-N-size-paper-FINAL.pdf

Dworkin, L. & Hyslop, A. (2019). *Screened out? How some states may limit the schools, and students, identified for support.* Alliance for Excellent Education. https://all4ed.org/screened-out-how-some-states-may-limit-the-schools-and-students-identified-for-support/

Gordon, N. (2017). *How state ESSA accountability plans can shine a statistically sound light on more students*. The Brookings Institution. https://www.brookings.edu/research/how-state-essa-accountability-plans-can-shine-a-statistically-sound-light-on-more-students/

Hough, H., Penner, E., & Witte, J. (2016). *Identity crisis: Multiple measures and the identification of schools under ESSA.* Policy Analysis for California Education. https://www.edpolicyinca.org/sites/default/files/PACE_PolicyMemo_1603.pdf

Rentner, D. S., Kober, N., & Braun M. (2019). *How states are responding to ESSA's evidence requirements for school improvement*. Center on Education Policy. https://www.cep-dc.org/displayDocument.cfm?DocumentID=1502

Sabia, R., & Cortiella, C. (2017). *Every Student Succeeds Act (ESSA) state plan review guide & advocacy tips*. National Down Syndrome Congress and The Advocacy Institute. https://www.advocacyinstitute.org/ESSA/ESSA.State.Plan.Review.Guide.Advocacy.Tips.June2017.pdf

Seastrom, M. (2017). *Best practices for determining subgroup size in accountability systems while protecting personally identifiable student information* (IES 2017-147). U.S. Department of Education, Institute of Education Sciences. Retrieved January 27, 2020, from https://nces.ed.gov/pubs2017/2017147.pdf

Walsh, E., Liu, A. Y., & Dotter, D. (2014). *Measuring Teacher and School Value Added in Oklahoma, 2012-2013 School Year*. Washington, DC: Mathematica Policy Research. https://www.mathematica.org/our-publications-and-findings/publications/measuring-teacher-and-school-value-added-in-oklahoma-20122013-school-year

Wasserstein, R., & Lazar, N. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. https://amstat.tandfonline.com/doi/full/10.1080/00031305.2016.1154108#.XxqnMp5KhPY

October 2020

This report was prepared for the Institute of Education Sciences (IES) under Contract ED-IES-17-C-0006 by the Regional Educational Laboratory Mid-Atlantic administered by Mathematica. 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.

This REL report is in the public domain. While permission to reprint this publication is not necessary, it should be cited as:

Forrow, L., & Kelly, K. (2020). *Why school accountability systems disproportionately identify middle schools' SWD subgroups for TSI* (REL 2021-049). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic. http://ies.ed.gov/ncee/edlabs.

This report is available on the Regional Educational Laboratory website at http://ies.ed.gov/ncee/edlabs.

Appendix A. Data and descriptive analysis methods

Data

Using publicly available data from State A's and State B's public data portals, we compiled a data set containing information on the number of counted students and indicator scores for each school in each state, separately for the all students group and students with disabilities (SWD) subgroup. Because the public-use data sets did not contain complete information about all relevant accountability indicators, we focused on the following indicators, using 2017/2018 school year information: English language arts (ELA) proficiency, math proficiency, ELA achievement growth, math achievement growth, regular attendance, and four-year graduation rates (for high schools only). We classified each school as elementary, middle, or high based on the grade levels represented in the school.[5] We also gathered statewide information on the number of students and average performance on the accountability indicators from the public data sets.

Public data files suppress information based on small samples to protect students' privacy. For example, if a school had only five SWDs, the public data set would not report the number of SWDs or the average SWD accountability indicator scores for that school. This practice limited the information available for analysis, sometimes substantially. However, the public use files provided adequate information to assess why schools' indicator scores were suppressed, which allowed us to gauge the sensitivity of our results to this suppression. Public use data files suppress indicator scores because the underlying sample size did not meet the state's minimum n-size requirement or because the school's average indicator score was so low that it would violate students' privacy to report it. Scores that are suppressed due to sample size would not be included in accountability calculations, so their absence from the data file does not bias our results. However, scores that are suppressed due to indicator score would be included in accountability calculations, and they would account for a moderate proportion of the suppressed ELA and math proficiency scores for SWD subgroups in State B. In the methods section we describe a sensitivity test that gauges the influence of these suppressed scores on our findings.

Methods

We focused our first set of descriptive analyses on the rates at which schools at different levels meet their state's minimum n-size requirements. Specifically, for each state and school level, we calculated the percentage of schools with enough tested SWDs to meet the minimum n-size requirement for at least one academic proficiency indicator.

In a second set of descriptive analyses, we compared the scores of all students and SWDs on each of the six accountability indicators examined for both states by school level to determine whether the relative performance of SWDs differed by school level across the indicators that contribute to the accountability system. To acknowledge that larger schools contribute more students to the statewide average than smaller schools, in this analysis we weighted each school's indicator scores by the number of students enrolled at that school in that subgroup. For example, we weighted the all students group's ELA proficiency score by the total number of students at the school, and we weighted the SWD subgroup's ELA proficiency score by the total number of SWDs

[5] For consistency, we used the same rules in both states. We considered a school an elementary school if it had students in grades K to 5 and no students above grade 6; a middle school if it had students in grades 7 and 8 but not below grade 5 or above grade 9; and a high school if it had students in grades 10 to 12 and no students below grade 8. For schools that included grades in more than one of the three levels, such as K to 8 schools, we classified them based on the majority of grades represented.

at the school.[6] Although ideally we would have weighted each indicator score by the number of students with scores for that indicator, these indicator-specific sample sizes were not consistently available in the public data sets. Instead, we chose to weight by the total number of students in each subgroup, because this weighting approach allows larger schools to contribute more to the statewide average than smaller schools, more closely approximating a direct student-level average.

When calculating average indicator scores at each school level in each state, we excluded suppressed indicator scores. As noted above in the Data section, if these indicator scores are suppressed because the school did not meet the minimum n-size for that indicator, this procedure appropriately mirrors states' accountability calculations. However, if the indicator scores are suppressed for privacy reasons and would actually be included in accountability calculations, omitting them could change our results. Diagnostics suggest that such rules account for a negligibly small proportion of suppressed indicator scores in State A, but they account for a moderately large proportions of suppressed ELA and math proficiency indicator scores for SWD subgroups in State B. Representatives from State B confirmed that these scores are largely suppressed because the SWD subgroup's proficiency rate was so low that it would violate students' privacy to report it; thus, as a sensitivity test, we imputed a school's average indicator score as 0 if the indicator score was suppressed but the corresponding sample size met the state's minimum n-size requirement. After imputing the suppressed indicator scores that met these criteria, we recalculated the average indicator score for each combination of state, school level, and subgroup. Then, we recalculated the performance gap between the all students group and the SWD subgroup for each combination of state and school level. The inclusion of these schools' subgroups' scores as 0 does not affect our conclusion (figure C.2 in appendix C).

[6] Public data sets from State B did not provide the number of SWDs at each school. The public data sets included the total number of students and the percentage of students who were SWDs. We calculated the number of SWDs at each school in State B as the floor of the product of these two values.

Appendix B. States' accountability rules

State A

We calculated a school's summative score as the weighted average of its raw performance indicator scores, using the weights provided in the state's Every Student Succeeds Act plan. We took separate weighted averages using performance indicator scores based on the overall student population and the SWD population to create summative scores for all students and SWDs.

Table B.1. Composite weights for State A by school level

Indicator	Elementary	Middle	High
English language arts proficiency	15	15	15
Math proficiency	15	15	15
English language arts growth	15	15	0
Math growth	15	15	0
Regular attendance rate	10	10	5
Four-year graduation rate	0	0	15

Source: Accountability system technical guidelines for State A.

State B

State B's accountability system is more complex, involving several exclusion criteria to filter out schools with inadequate data and several transformations to standardize the indicator distributions. We describe the system step-by-step here:

1. Assign each school to a category, called its configuration, based on the number of accountability indicators with non-missing data. These categories group schools with similar numbers of non-missing accountability indicator scores so that, when ranking schools' accountability scores to make CSI and TSI determinations, each school is compared to other schools with similar information. We adapted these criteria slightly to account for the smaller set of indicators included in the simulation.
    a. For elementary and middle schools, we assigned a school with a non-missing score for at least three of ELA proficiency, math proficiency, ELA growth, and math growth the "ES/MS" configuration; we exempted schools with fewer than three non-missing scores on these indicators from the accountability calculation.
    b. For high schools, we assigned a school with a non-missing score for at least two of ELA proficiency, math proficiency, and the four-year graduation rate the "HS" configuration; we excluded schools with fewer than two non-missing scores on these indicators from the accountability calculation.
    c. Because subgroups are generally smaller than the school overall, a given subgroup may have fewer non-missing accountability indicators than the school as a whole. For that reason, we separately assigned a configuration to each school for its SWD population, based on the number of indicators with non-missing scores for SWDs.
2. Calculate an All Student Standard Score for each indicator. This calculation is shown below, using math proficiency as an example.
    a. Each school has a score for math proficiency. For example, 78 percent of tested students are proficient.
    b. Each school's raw score is transformed into a *z*-score following the distribution of all other schools with the same configuration.

    c. If School X has a math proficiency *z*-score of 0.5, its All Student Standard Score for math is 0.5.

3. Calculate a Subgroup Standard Score for each indicator following the same steps.
    a. Each school has a math proficiency score calculated among subgroups of students, such as SWDs. For example, 25 percent of tested SWDs at School X are proficient.
    b. Each school's raw subgroup score is transformed into a *z*-score using the distribution of scores for the same subgroup across all schools with the same configuration.
    c. Supposing that 25 percent of SWDs at School X are proficient, we might calculate a Subgroup Standard Score (*z*-score) of -0.2 for SWDs when we compare the performance of School X's SWDs with the performance of SWDs at other schools with the same configuration.
4. Calculate the Average Standard Score as the average of the All Student Standard Score and Subgroup Standard Scores for each indicator. For example, for math proficiency:
    a. Take the average of the school's All Student Standard Score and its Subgroup Standard Scores; here we suppose that SWDs are the only subgroup at the school.
    b. For School X, the Average Standard Score for math proficiency is $\frac{0.5+(-0.2)}{2} = 0.15.$
5. Calculate the Indicator Score for math proficiency by converting the Average Standard Score to a percentile rank based on the distribution of Average Standard Scores across schools with the same configuration. Suppose School X receives a math proficiency Indicator Score of 65.
6. Calculate the Summative Score as the weighted average of Indicator Scores for each performance indicator, using the weights specified in the state Every Student Succeeds Act plan (table B.4).
7. Take the Summative Score calculated in Step 6 as the school's overall Summative Score, used for CSI determinations. Each school also receives a Summative Score for each subgroup; these scores are used for TSI determinations.
    a. The simulation included data on a single subgroup: SWDs. The Indicator Score for SWDs for each indicator is simply the school's *z*-score for that measure from step 3 above.
    b. Calculate the Indicator Score for each performance indicator by finding the percentile rank of each subgroup's Standard Score relative to the distribution of Standard Scores for that indicator for that subgroup across all other schools with the same configuration, using the configurations defined for the SWD subgroup in step 1.c.
    c. Calculate the subgroup's Summative Score as a weighted average of Indicator Scores using the same indicator weights as in the all students group.

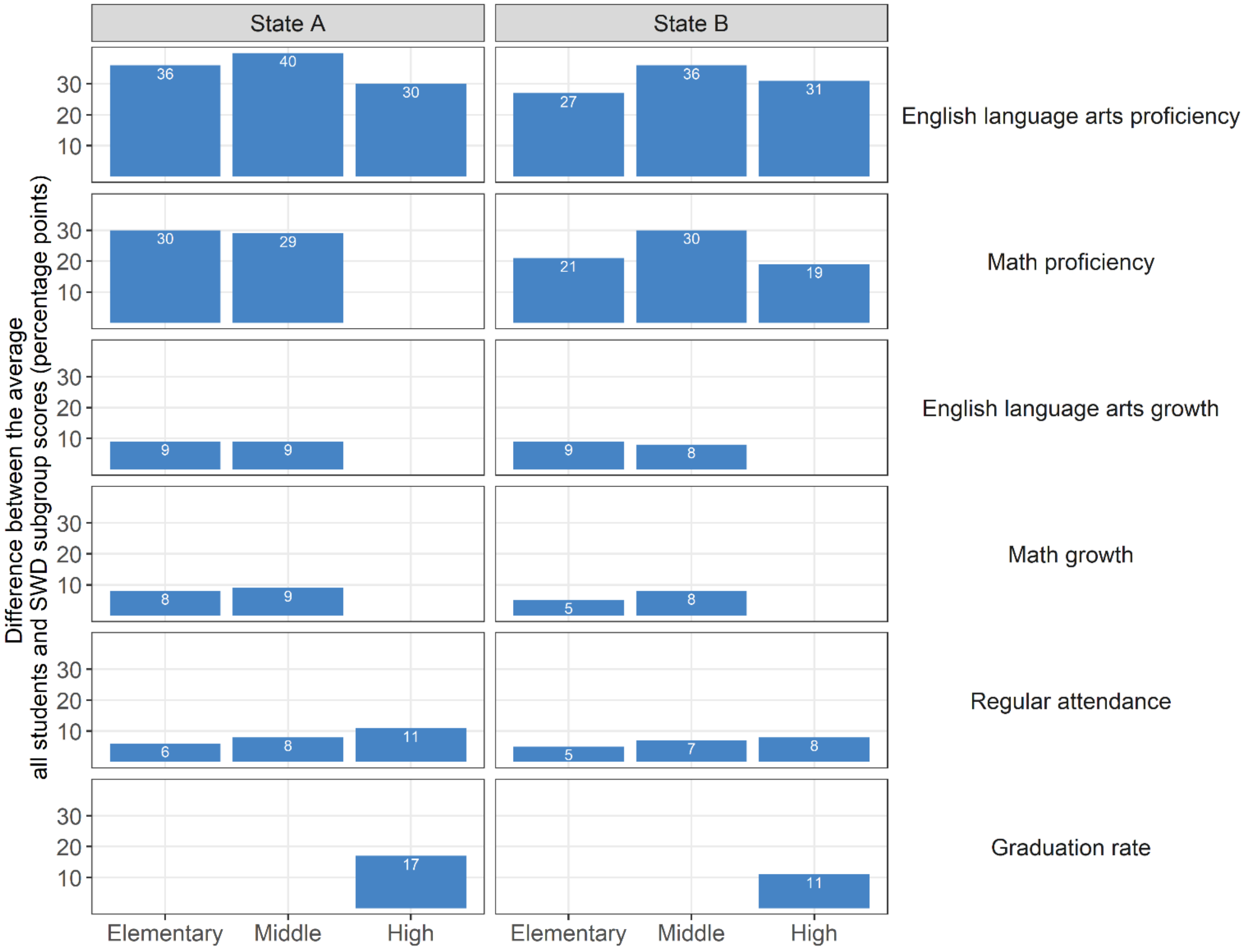
Table B.2. Composite weights for State B by school level

Indicator	Elementary	Middle	High
English language arts proficiency	17.5	17.5	17.5
Math proficiency	17.5	17.5	17.5
English language arts growth	25	25	0
Math growth	25	25	0
Regular attendance rate	15	15	15
Four-year graduation rate	0	0	25

Source: Accountability system technical guidelines for State B

Appendix C. Supplementary analyses

Figure C.1. Differences between all students and SWD subgroup scores for all study indicators



SWD is students with disabilities.

Note: The height of each bar is calculated as the difference between the average indicator score for all students minus the average indicator score for the SWD subgroup across schools meeting the minimum n-size criterion based on all indicators. The average math proficiency score for the State A high school SWD subgroup was redacted in the original data, so no value is available for this score. ELA and math growth are not accountability indicators for high schools and graduation rate is not an accountability indicator for elementary or middle schools, so no values are available for those scores.

Source: Publicly available data from the departments of education of States A and B.

Table C.1 Average indicator scores by state, indicator, school level, and subgroup

State	Indicator	Elementary			Middle			High		
		All students	*SWDs*	*Difference*	*All students*	*SWDs*	*Difference*	*All students*	*SWDs*	*Difference*
State A	English language arts proficiency	57%	20%	36%	50%	10%	40%	50%	20%	30%
State A	Math proficiency	50%	21%	30%	36%	7%	29%	30%		
State A	English language arts growth	70%	60%	9%	56%	47%	9%			
State A	Math growth	67%	59%	8%	48%	39%	9%			
State A	Regular attendance	89%	83%	6%	83%	75%	8%	78%	68%	11%
State A	Graduation rate							87%	70%	17%
State B	English language arts proficiency	55%	29%	27%	62%	25%	36%	54%	22%	31%
State B	Math proficiency	49%	28%	21%	49%	19%	30%	39%	19%	19%
State B	English language arts growth	51%	42%	9%	49%	41%	8%			
State B	Math growth	51%	46%	5%	49%	41%	8%			
State B	Regular attendance	91%	86%	5%	91%	85%	7%	85%	78%	8%
State B	Graduation rate							92%	81%	11%

SWD is students with disabilities.

Note: The average math proficiency score for the State A high school SWD subgroup was redacted in the original data, so no value is available for this score. ELA and math growth are not accountability indicators for high schools and graduation rate is not an accountability indicator for elementary or middle schools, so no values are available for those scores.

Source: Publicly available data from the departments of education of States A and B.

Table C.2 Percentage of indicator scores suppressed for privacy reasons, by state, school level, indicator, and subgroup

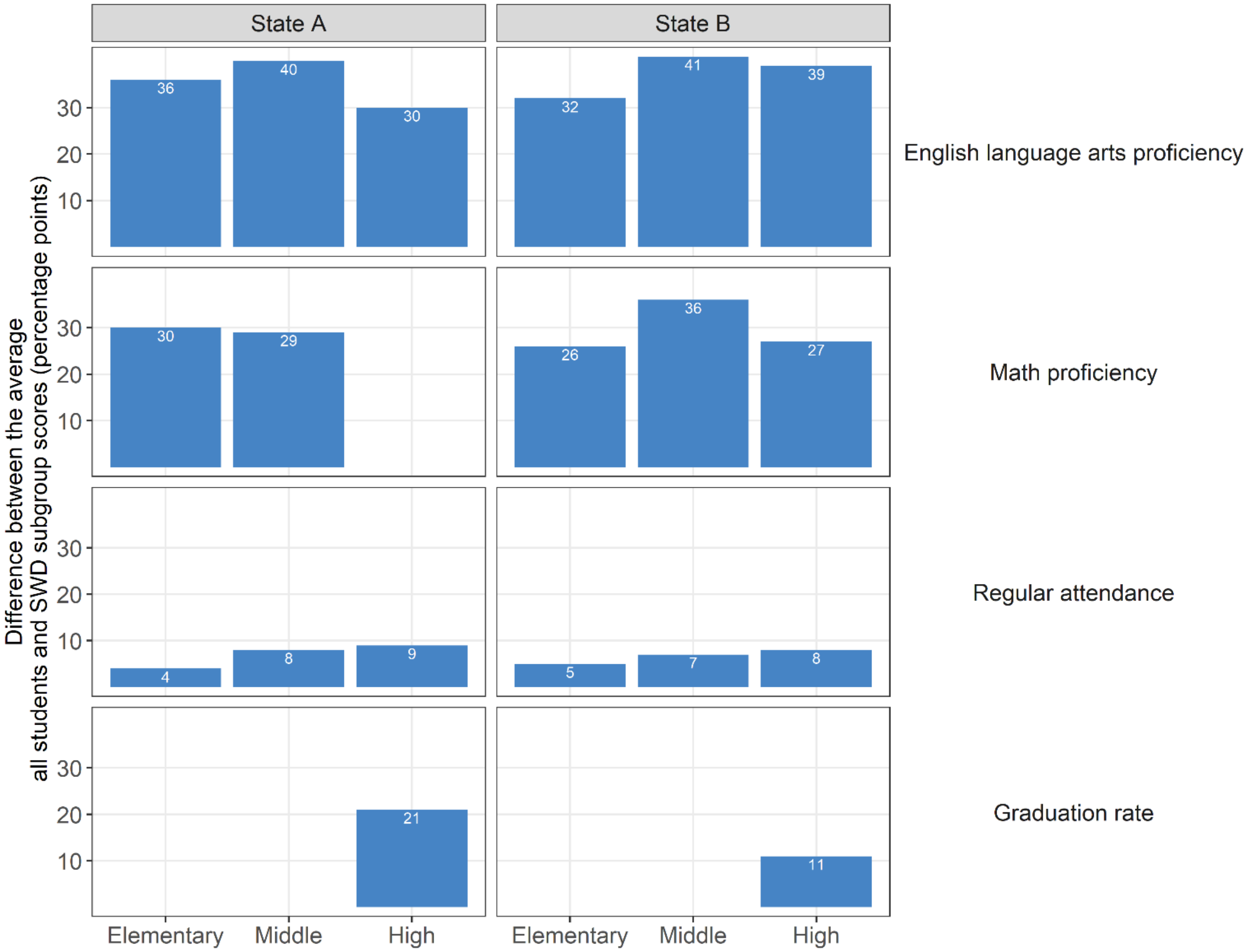
State	Indicator	School level	Subgroup	Percent of all schools	Percent below minimum n-size
State A	Regular attendance	Elementary	All students	0.0	na
			SWDs	9.5	na
		Middle	All students	0.0	na
			SWDs	2.9	na
		High	All students	0.0	na
			SWDs	11.5	na
	4-year graduation rate	High	All students	11.5	na
			SWDs	17.3	na
State B	ELA proficiency	Elementary	All students	0.5	0.0
			SWDs	15.6	11.6
		Middle	All students	0.0	na
			SWDs	17.8	3.2
		High	All students	3.5	0.0
			SWDs	31.2	3.5
	Math proficiency	Elementary	All students	1.9	0.0
			SWDs	18.5	8.5
		Middle	All students	2.5	0.0
			SWDs	35.7	1.6
		High	All students	14.7	5.9
			SWDs	54.5	5.6

ELA is English language arts. na is not applicable. SWD is students with disabilities.

Note: The table counts the number of schools at each school level in each state where the score for a given indicator was suppressed for privacy reasons, rather than because the school did not meet the minimum n-size requirement for that indicator. A state may suppress indicator scores for privacy reasons if they feel that displaying the data could lead to the inference of an individual student's data. For example, State B suppresses proficiency indicator data when indicator scores are below 10% so that schools or student groups with 0 percent proficiency rates cannot be identified. The percent of all schools is calculated as the number of schools with an indicator score suppressed for privacy reasons divided by the number of schools at that level in that state. For example, in the first row, we divide the number of elementary schools with suppressed regular attendance data by the total number of elementary schools in State A. For State B, we calculated the percent below minimum n-size as the percentage of schools with data suppressed for privacy reasons where the school would not have met the state's minimum n-size requirement for accountability. In State B, the minimum n-size required to report an indicator score in public files is lower than the minimum n-size required to include it in accountability calculations.

Source: Publicly available data from the departments of education of States A and B.

Figure C.2. Differences between all students and SWD subgroup academic proficiency scores, with imputation of indicator scores suppressed for privacy reasons



SWD is students with disabilities.

Note: The height of each bar is calculated as the difference between the average indicator score for all students minus the average indicator score for the SWD subgroup across schools meeting the minimum n-size criterion based on all indicators. Indicator scores that are suppressed for privacy reasons are imputed as 0; no indicator scores for ELA or math growth were suppressed for privacy reasons in either state, so we do not report those indicators here. The average math proficiency score for the State A high school SWD subgroup was redacted in the original data, so no value is available for this score. Graduation rate is not an accountability indicator for elementary or middle schools, so no values are available for those scores.

Source: Publicly available data from the departments of education of States A and B.