Economic Studies
at BROOKINGS

Evidence Speaks Reports, Vol 2, #17
June 29, 2017

How state ESSA accountability plans can shine a statistically sound light on more students

Nora Gordon

Executive Summary

The subgroup requirements for accountability in the No Child Left Behind Act (NCLB) were designed to reveal underperformance of disadvantaged groups that could otherwise be hidden in aggregate averages. Both NCLB and its successor, the Every Student Succeeds Act (ESSA), left the choice of minimum subgroup size at the school level (n-size) for accountability purposes to the states. A smaller n-size is more likely to include students from subgroups that are underrepresented at a particular school in considering accountability for that school, but decreases the statistical reliability of the estimate of how students in that subgroup are performing and can raise privacy concerns. Equity-oriented groups that want as many students from disadvantaged groups as possible included in the accountability system, including the Alliance for Excellent Education and the Education Trust, have advocated for states to adopt a minimum n-size of 10, whereas since revoked Obama-era accountability regulations allowed states to choose any n-size up to 30.

As the 34 states currently finalizing their ESSA accountability plans for the federal September deadline strive for comprehensive, context-specific strategies to cover more students and schools while maintaining statistical reliability, they should view minimum n-size as just one piece of these strategies. States must consider n-size alongside how they permit schools to combine data over grade levels, school years, and/or groups of students—strategies many states have been using under NCLB waivers and that first-round states have included in their ESSA plans.

I use national school-level enrollment data by race/ethnicity to show how many students in different subgroups are covered under different pooling approaches. Pooling data across years and grades will include most students in accountability systems, but for lower enrollment populations, pooling across racial/ethnic groups may provide an opportunity to include students in accountability systems in cases where subgroup size is otherwise too small. Each state should consider the demographic composition of its own districts in making its policy choices, not only about minimum n-size, but also about how districts can combine data to increase the number of students included.

A large body of research makes a convincing case that the design of accountability systems influences how schools respond to them.[1] The subgroup requirements for accountability in NCLB were designed to reveal underperformance of disadvantaged groups that could otherwise be hidden—inadvertently or not—in aggregate averages. However, the potential benefits of transparency from disaggregation—and corresponding decreases in sample size—come with trade-offs: smaller samples have more statistical noise and, if sufficiently small, may raise privacy concerns.[2]

This piece assesses common ways states have combined data over the subgroup-grade level-year level to increase effective sample size and statistical reliability. I use national data to show how many students would be excluded from accountability measures at different n-sizes and how other combining data across grade levels or years within a school can include more students in the accountability process. These pooling methods allow states to retain the benefits of privacy and statistical reliability of larger minimum subgroup reporting sizes while including more schools and students in the system.

How subgroups work

States use subgroups for two purposes, with potentially two different minimum subgroup sizes, or n-sizes: reporting (school report cards available to the public online) and federal accountability (used in state calculations to determine which schools fall into particular categories under ESSA). Under federal law, states are responsible for determining minimum n-sizes. States can and sometimes do choose different cutoffs for n-size for reporting and accountability.

This piece relates solely to n-size for accountability. If n is too small, statistical reliability is at risk; if n is too big, too few schools and students are held accountable, as those with subgroup enrollments less than n do not participate in the accountability system.

Consider an elementary school with grades K-5. It is required to report whatever metrics its state chooses not only for all its tested grades (3-5), but also for a number of distinct "subgroups" including those defined by race/ethnicity, as long as there are more students in each subgroup than the minimum n-size the state has chosen. For example, assume the school has 20 Hispanic third-grade students, 15 Hispanic fourth graders, and 9 Hispanic fifth graders, and is in a state that set its n-size at 30. Without combining data in some way, such as across grade levels or school years, Hispanic students won't be included as a separate group in the state's accountability system for that school. If the state's n-size were 20, only the third grade Hispanic students at the school would be included; if n-size were 10, both third and fourth graders separately would be included. But if the school pooled data from all tested grades, it would have 44 Hispanic students in grades 3-5 collectively, well over the n-size of 30.

Current policy and advocacy related to subgroup size

The ESSA did not specify any maximum cutoff for minimum n-size in state plans, though it did mandate that the Institute of Education Sciences (IES) issue a technical report to provide states with background on the topic. That report highlights tradeoffs between statistical reliability and privacy versus covering more students, and does not define an optimal n-size.[3] In its (now rescinded) accountability regulations, the Obama administration ultimately told states they had to explain any minimum n-sizes greater than 30, arguing that larger n-sizes are not necessary for either statistical reliability or privacy protection.[4] Though the regulations are no longer the law, among the states that have turned in their ESSA plans to date, none have n-sizes greater than 30.[5] The current U.S. Department of Education guidance on accountability under ESSA does not define any cutoff for minimum n-size.[6]

A number of civil rights and education reform groups advocated for n-size of 10 in the ESSA accountability regulations. These efforts are driven by the desire to hold schools accountable for the performance of *all* their students—a hugely important goal that history does not allow us to take for granted. But the motivation does not need to map to the narrow focus on n-size at a particular grade, without considering pooling of data. Indeed, advocates should care about the implications of increasing statistical volatility when using a small sample for accountability purposes.

The Alliance for Excellent Education points to the 2010 report from IES' National Center for Education Statistics (NCES) on state longitudinal data systems as support for choosing an optimal n-size of 10.[7] The NCES report, however, is about subgroup size for reporting purposes, not accountability.[8] It correspondingly focuses on privacy concerns, rather than ensuring statistical reliability for high stakes policy

decisions, arguing that n-size of 10 protects student privacy. In practice, the n–size discussion is now about the range of n=10 to 30, so the real issue here is statistical reliability rather than privacy.

As sample size shrinks, the chances rise that a few individual children influence the school's accountability rating—either positively or negatively—in a way that has nothing to do with how well the school serves students in that subgroup. And while accountability metrics that rely on gains are statistically preferable to proficiency ones, gains are even more subject to volatility when samples are small. Because real stakes are attached to these accountability ratings, states should tread carefully.

Options for covering more students and schools

Although the most visible advocacy efforts to cover more students from traditionally disadvantaged groups have focused on getting states to decrease n-size, other strategies, including data averaging and the judicious use of super subgroups, can serve that purpose while mitigating the concerns over statistical reliability and privacy prompted by very low n-sizes.

Data averaging pools students across grade levels at the school-year level, and/or across years at the school-grade level, increasing the number of observations and, consequently, the chance of getting to the minimum n-size. For example, an elementary school with grades K-5, and therefore tested grades 3-5, could average outcomes for Hispanic students in grades 3, 4, and 5, summing the number of Hispanics in each of the three grades to arrive at a higher n-size. It could also average those outcomes from the current school year with those of the previous year or two: for example, scores of third-grade Hispanic students in spring 2017 averaged with those of third-grade Hispanic students in spring 2016, or 2016 and 2015. Current federal guidance explicitly permits states to combine data across grades within a school or across school years.[9]

Different ways of combining data come with different trade-offs: pooling across years does not hold schools accountable for individual year-to-year changes, while pooling across grades masks potential differences across grade levels within a school. Either way has the benefit of including schools and students in accountability systems when they would otherwise be exempt due to sample size below n, or included in a system with a small but statistically unreliable n.

Super subgroups can be formed by aggregating data in a variety of ways within a grade level and school year. Under NCLB waivers, many states used this general concept, with a range of custom designs.[10] Federal guidance for accountability under ESSA touches on this topic, and is difficult to interpret definitively. It notes that states must include all schools in accountability systems and may need to use alternate methodologies to include some schools based on their specific contexts, if they remain uncovered after they have combined data across grades and years.[11] The same document prohibits states from combining "major racial and ethnic subgroups…into a…'super-subgroup,' as a substitute for considering student data in each of the major racial and ethnic groups separately (emphasis added)."[12] The guidance does *not* explicitly prohibit aggregation of data across racial and ethnic subgroups in cases in which data for those disaggregated categories would be impossible to report due to sample sizes falling below the state's specified n-size.

Advocates seeking transparency for individual racial/ethnic subgroups of students have been vocal in their opposition to the "super subgroup" approach. As the data below show, however, for some low enrollment groups, this approach can increase coverage in ways that data averaging cannot.

Examples from Oregon's state plan

Oregon's plan provides two examples of how states can hold districts accountable for a greater share of students in "underserved" groups for any given n-size choice.[13] Its state plan is worth reading for how it walks through its chosen strategy, demonstrating the impact of design elements on the share of students included in accountability plans.

- Disaggregated subgroups by race/ethnicity plus "combined underserved race/ethnicity" student group. Wherever n-size (20 in Oregon's plan) permits, school-level accountability will use disaggregated data for each racial/ethnic subgroup. In addition, the state will continue to use its "combined underserved race/ethnicity," combining the four racial/ethnic subgroups with achievement gaps in Oregon.

  This super subgroup will be used for accountability *only in cases where no disaggregated subgroup*

of these students meets the minimum n-size. That is, the super subgroup in the Oregon plan is only a substitute for no data, not for disaggregated data. It differs in that regard from Delaware: the U.S. Department of Education informed Delaware this month that its plan must be revised because it used a super subgroup *without* including disaggregated subgroups.[14]

- Data averaging over multiple years. Oregon requires districts to report three-year averages in addition to one year. Again, this measure is only used for accountability purposes when there is insufficient sample size to use the current year measure, rather than as a default.

How these strategies would affect coverage nationally

To show how such strategies increase coverage, I turn to the NCES Common Core of Data's Public School Universe for 2014-15.[15] Among those elementary schools reporting enrollment by grade level, I further limit the sample to those with students enrolled in grades 3, 4, and 5, but not in higher grades. This yields a sample of 26,710 schools nationally.

For n-size of 10, 20, and 30 (the three most common n-sizes in state plans submitted in the first round), I calculate what shares of Black and Hispanic students would be covered in state accountability systems under a set of four distinct regimes:

1. Most disaggregated: This column uses the n-size for each grade level, racial/ethnic subgroup, and school year, with no further aggregation.
2. Grade-span reporting: If enrollment levels are relatively constant by group over grades, this strategy essentially multiplies the number of subgroup observations by the number of covered grades in a school. This sample is constructed to contain solely tested grades 3, 4, and 5 for each school, so would multiply the most disaggregated subgroup count by three.
3. A combined underserved subgroup similar to Oregon's: aggregating American Indian or Alaskan Native, Black or African American, Hispanic/Latino, and Native Hawaiian or other Pacific Islander students within each grade level. (The extent to which this increases sample size depends on the demographic composition of the school.)
4. Two-year data averaging: using two school years' worth of data on the racial/ethnic subgroup for that grade level, so drawing on two cohorts of students. Again assuming a constant distribution of students by race/ethnicity in the school over the two-year time period, this increases the number of students in each grade-level racial/ethnic subgroup by a factor of the number of years averaged, here two. A state using three-year data averaging would increase subgroup sample size similarly to using grade-span reporting over three grade levels.

In practice, states may and sometimes do combine strategies. I examine each separately in the tables below, using actual data from NCES Common Core of Data's Public School Universe in 2014-15.

Table 1. Percent Black third graders nationally unaccounted for, by subgroup design

n-size	Most disaggregated	Grade span reporting (3-5)	Combined underserved subgroup	Two-year data averaging
n=10	11.2%	3.2%	2.5%	4.6%
n=20	25.0%	7.5%	8.2%	11.2%
n=30	38.9%	12.4%	16.0%	18.2%

The most disaggregated column in Table 1 shows how increasing n-size leaves more students unaccounted for: with n-size of 10, only 11 percent of Black students in third through fifth grade would be left out of accountability systems, whereas with n-size of 30—and no other strategy to pool data—39 percent would be left out. This is the argument behind the Education Trust's push for n-size of 10.

If states choose to pool data in other ways, however, they can support larger minimum n-size and its statistical and privacy benefits, with a much smaller hit to coverage than choosing the most disaggregated policy with n-size of 30. A state where policymakers and stakeholders have strong preferences for statistical reliability could choose to combine data across grade span, years, or subgroups, depending on the policy goals. Is it more valuable to know about third grade performance, across multiple years, or third, fourth, and fifth grade performance in a single school year? This answer should help inform the measurement strategy.

State-level variation

How many students would be left uncovered using any of these strategies depends on local demographics and segregation patterns. Table 2 shows how the share of Hispanic third graders not covered by a subgroup varies by subgroup definition strategy across the states.

Table 2. Variation in how subgroup design affects coverage of third graders, by race/ethnicity

	Most disaggregated	Grade span reporting (3-5)	Combined underserved group	Two-year data averaging
Black/ African American	38.9%	12.4%	16.0%	18.2%
Hispanic	26.5%	8.6%	15.1%	12.5%
American Indian/ Alaskan Native	76.1%	55.5%	37.0%	60.8%
Hawaiian Native/ Pacific Islander	76.8%	51.7%	15.2%	58.9%

Table 2 shows that different strategies have different effects for representation for different subgroups, holding the minimum n-size constant at 30. Aggregating across grade span or year is more effective for higher enrollment subgroups—Black and Hispanic students. For smaller American Indian/ Alaskan Native and Hawaiian Native/Pacific Islander subgroups, the majority of students in the subgroup remain uncovered if only students in that subgroup are pooled: the "super subgroup" strategy of aggregating across racial/ethnic groups is the only way to account for most students in these groups, although their data are not identifiable at the subgroup level.

What this means for states still finalizing their plans

States still finalizing their plans should know that n-size of 10 is not the only way to cover most students—and that the path to statistical reliability doesn't require excluding lots of students. Pooling data across years and grades will include most students in accountability systems, but for lower enrollment populations, pooling across racial/ethnic groups may provide an opportunity to include students in accountability systems in cases where subgroup size is otherwise too small. In the national data represented in Table 2, this would include American Indians/Alaskan Natives and Native Hawaiians/Pacific Islanders, but each state should inform its policy choice based on the demographic composition of its own districts. The analysis in the Oregon state plan provides an excellent example of how states can explore the policy trade-offs relevant to their own contexts.

---

[1] Loeb, Susanna, and David Figlio. 2011. "School accountability". In Eric A. Hanushek, Stephen Machin, and Ludger Woessmann (Eds.), *Handbook of the Economics of Education*, Vol. 3 (pp. 383-423). North Holland.

[2] Kane, T. J., Staiger, D. O., Grissmer, D., and Ladd, H. F. 2002. Volatility in School Test Scores: Implications for Test-Based Accountability Systems. *Brookings Papers on Education Policy*, No. 5, pp. 235-283. Also see: Kane, Thomas, and Douglas O. Staiger. 2002. "The Promise and Pitfalls of Using Imprecise School Accountability Measures." *Journal of Economic Perspectives*, 16 (4): 91–114.

[3] Seastrom, Marilyn. 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. Washington, DC. Retrieved from: https://nces.ed.gov/pubs2017/2017147.pdf.

[4] Elementary and Secondary Education Act of 1965, as Amended by the Every Student Succeeds Act—Accountability and State Plans, Vol. 81, No. 229 Fed. Reg. (November 29, 2016) (to be codified at 34 CFR Parts 200 and 299). https://www.gpo.gov/fdsys/pkg/FR-2016-11-29/pdf/2016-27985.pdf.

[5] "Key Takeaways: State Accountability Plans Under ESSA." *Education Week*. 12 April 2017. Retrieved from: http://www.edweek.org/ew/section/multimedia/key-takeaways-state-essa-plans.html#detailed-overview.

[6] Accountability under Title I, Part A of the ESEA: Frequently Asked Questions. US Department of Education. January 2017. Retrieved from: https://www2.ed.gov/programs/titleiparta / eseatitleiaccountabilityfaqs.pdf.

[7] Cardichon, Jessica. 2016. “Ensuring Equity in ESSA: The Role of N-Size in Subgroup Accountability.” *Alliance for Excellent Education*. Retrieved from: http://all4ed.org/wp-content/uploads/2016/06/NSize.pdf.

[8] National Center for Education Statistics. 2010. “Statistical Methods for Protecting Personally Identifiable Information in Aggregate Reporting.” SLDS Technical Brief. https://nces.ed.gov/pubs2011/2011603.pdf.

[9] Accountability under Title I, Part A of the ESEA: Frequently Asked Questions. US Department of Education. January 2017. See C-11.

[10] Ujifusa, Andrew. “ESSA Means the End of How These States Use 'Super Subgroups' for Accountability.” *Education Week*. 21 June 2016. Retrieved from: http://blogs.edweek.org/edweek/campaign-k-12/2016/06/ESSA_super_subgroups_accountability_changes.html.

[11] Accountability under Title I, Part A of the ESEA: Frequently Asked Questions. US Department of Education. January 2017. See C-26.

[12] Ibid, see E-8.

[13] Huckaby, Dawne. 2017. “Oregon’s Consolidated State Plan under the Every Student Succeeds Act.” Oregon Department of Education. Retrieved from: https://www2.ed.gov/admins/lead/account/stateplan17/orcsa2017.pdf.

[14] Botel, J. (2017, June 13). [Letter to Susan Bunting]. Delaware Department of Education, Dover, Delware. Retrieved from: https://www2.ed.gov/admins/lead/account/stateplan17/deprelimdetermltr.pdf.

[15] 2014–15 Common Core of Data (CCD) Universe Files (NCES 2016-077). U.S. Department of Education. Washington, DC: National Center for Education Statistics.