CADRE

AUGUST 2025

Investigating the Relationship Between Sample Size and Reliability of Aggregate Test Score Measures in Colorado Schools

Benjamin R. Shear, Erik Whitfield, Kaitlin Nath

A report prepared by the Center for Assessment, Design, Research and Evaluation (CADRE) at the CU Boulder School of Education.

School of Education
UNIVERSITY OF COLORADO **BOULDER**

About CADRE

The Center for Assessment, Design, Research and Evaluation (CADRE) is housed in the School of Education at the University of Colorado Boulder. The mission of CADRE is to produce generalizable knowledge that improves the ability to assess student learning and to evaluate programs and methods that may have an effect on this learning. Projects undertaken by CADRE staff represent a collaboration with the ongoing activities in the School of Education, the University, and the broader national and international community of scholars and stakeholders involved in educational assessment and evaluation.

Acknowledgements

The authors acknowledge helpful feedback on earlier drafts of this report from Elena Diaz-Bilello at CADRE and Marie Huchton and Dan Mangan at CDE. All views expressed in the report and any remaining errors are the responsibility of the authors.

Suggested Citation

Shear, B. R., Whitfield, E., & Nath, K. (2025). Investigating the relationship between sample size and reliability of aggregate test score measures in Colorado schools. Boulder, CO: The Center for Assessment, Design, Research and Evaluation (CADRE), University of Colorado Boulder.

Please direct any questions about this report to:

benjamin.shear@colorado.edu

Table of Contents

Introduction

Determining minimum sample size requirements for reporting results from state accountability testing continues to be a complicated statistical and policy decision. The Every Student Succeeds Act (ESSA) legislation requires states to design assessment systems that “enable results to be disaggregated within each State, local education agency, and school” by demographic disaggregated groups (Every Student Succeeds Act, 2015 Sec. 1111(b)(2)(B)(xi)). However, the legislation notes that “disaggregation shall not be required in the case of a State, local educational agency, or a school in which the number of students in a disaggregated group is insufficient to yield statistically reliable information or the results would reveal personally identifiable information about an individual student” (Every Student Succeeds Act, 2015 Sec. 1111(b)(2)(B)(xi)). States balance these requirements by setting a “minimum n-size,” which indicates the minimum sample size needed for test score results from schools, districts, and disaggregated groups to be reported and used for school accountability (Sec. 1111(c)(3)).

Little consistency exists across states in determining minimum sample sizes that adhere to these requirements (Alliance for Excellent Education, 2018). An Institute of Education Sciences (IES) report providing guidance for setting minimum n-size requirements notes that “Choosing a minimum n-size is complex and involves important and difficult tradeoffs” (Seastrom, 2017, p. iv), while also noting that “ESSA prohibits IES from recommending any specific minimum number of students in a disaggregated group” (p. iv). As a result, states are responsible for determining the minimum n-size thresholds considered sufficient to yield “statistically reliable information” and to protect individual student privacy when reporting state assessment data for accountability ratings or other purposes. Minimum n-size thresholds for reporting and accountability range from 5 to 30 students across states and sometimes vary for different data sources within states (Alliance for Excellent Education, 2018).

Setting a minimum n-size requires balancing two competing interests. First, privacy and statistical reliability concerns tend to favor *higher* minimum n-sizes. With higher minimum n-sizes there is less chance that private information could be identified for individual students. Results reported for larger groups also tend to be more reliable and consistent over time because they are less influenced by sampling error and other chance variability. On the other hand, transparency and equity goals of ESSA tend to favor lower minimum n-sizes. With lower minimum n-sizes data will be reported for a greater number of disaggregated groups and schools, thus increasing transparency of the system.

Purpose

The *1241: Accountability, Accreditation, Student Performance, and Resource Inequity Task Force* (henceforth the "1241 Task Force"), appointed by the Colorado General Assembly in 2023, produced a final report with recommendations in November 2024[1]. The report made two recommendations related to minimum n-size requirements:[2]

- Lower student count thresholds for accountability calculations and reporting (Recommendation 1 in the report).

- Explore best practices and monitor the accountability system to identify and reduce issues of volatility that impact schools and districts with small student populations (Recommendation 4 in the report).

The present study investigates two questions using Colorado test score data for Elementary and Middle Schools that are relevant for informing implementation of these recommendations. These questions are:

1. How would the number of schools with reportable achievement and growth data be impacted by changes to the minimum n-size requirements?

2. How would the reliability of school-level achievement and growth results be impacted by changes to the minimum n-size requirements?

We address the first question by providing descriptive analyses of student sample sizes across Colorado schools, both overall and by relevant demographic disaggregated groups. The second question requires defining "reliability" more formally. To address the second question, we provide a formal definition of reliability and investigate how reliability would be impacted by changing sample size requirements. We also discuss the extent to which this formal definition of reliability is appropriate in the context of determining minimum n-size requirements, which is ultimately a policy decision, not a purely statistical decision.

[1] Final report at: https://www.cde.state.co.us/accountability/accountability-task-force

[2] These are recommendations #1 and #4 in the report. Recommendations 2 and 3 also are relevant to sample size requirements because they recommend combining or re-defining the disaggregated groups for which results are reported. However, we do not directly address those recommendations here, as the sample size considerations discussed here would be applicable to alternative definitions of student disaggregated groups.

Background

Current Colorado Minimum N-Size Thresholds

In Colorado there are two minimum n-size thresholds used to determine reporting of test score results, one for average test scores ("achievement") and one for aggregate student growth percentiles ("growth"). The minimum n-size for reporting achievement results is n=16 while the minimum n-size for reporting growth results is n=20. There is little formal documentation on the process to establish these requirements, but the requirements have been in place since 2010. In Colorado there have been at least two primary concerns with current minimum n-size requirements from school and district stakeholders, both primarily impacting smaller schools and districts. These concerns are cited in the Task Force Report as motivations for the recommendations listed above.

First, in small schools where sample sizes are at or just above the minimum n-size threshold, test score results tend to be highly variable from year to year. This occurs because with small sample sizes a small number of students can have a meaningful impact on aggregate results. Rating schools based on small samples of students may be considered less fair, because the accountability ratings and associated consequences are potentially influenced by inconsistent student-level factors outside the influence of schools or districts.

Second, and in contrast to the first concern, when a school has fewer student test scores than the minimum n-size, it is not possible to construct required accountability ratings and schools receive the state's "insufficient state data" (ISD) rating. When this occurs, there is too little information about student performance to produce an accountability rating and the school may not be eligible for, or if eligible, receive lower funding priority for Colorado Department of Education (CDE) school support systems. This lack of reporting also undermines transparency and equity goals that rely on reporting data for all student disaggregated groups.[3]

Aggregate Achievement and Growth Measures in Colorado School Accountability

Colorado uses aggregate school and district test score results to compute school and district ratings as part of the School Performance Framework (SPF) and District Performance Framework (DPF) scores, which we refer to as "Framework scores." The achievement component of Framework scores is calculated based on average test scores on CMAS (in grades 3-8) or the PSAT (in grades 9-10).[4] We refer to the average test scores in a school or district as the "achievement" measure. The growth component is calculated based on median student growth percentiles (SGP; Betebenner, 2009) based on CMAS (in grades 3-8) or the

[3] In some cases, CDE will aggregate the most recent three years of data for a school to increase sample sizes for accountability reporting. The aggregation of data across multiple years raises separate fairness concerns, as it results in some schools being evaluated on the most recent year of student performance, while others are evaluated based, in part, on student performance from as much as three years ago. The number of ISD ratings increased during the COVID-19 pandemic, which saw a decline in school enrollments and lowered test participation rates, thus pushing more schools either below or very near the minimum n-size thresholds.

[4] Although all Colorado 11thgraders take the SAT, these scores are included in the "postsecondary and workforce readiness" (PWR) indicator rather than the achievement SPF indicator.

PSAT/SAT (in grades 9-11). Although Colorado uses the median SGP, we analyze results for the arithmetic mean SGP (the mean SGP) because the statistical properties of the mean are more straightforward to analyze. Prior research has shown that mean SGP tend to have less sampling error and be more reliable than median SGP (Castellano & McCaffrey, 2017). Thus, we expect the results here based on the mean SGP will provide an upper bound on the reliability of median SGP.

To calculate SPF scores, school achievement and growth scores are assigned point values based on the distribution of results across schools and districts. Table 1 shows how average test scores and median SGPs are converted to points for Framework scoring. These scoring calculations are carried out for test score results based on all students in the school or district and for each of the following four disaggregated groups: minority students, students eligible for free- or reduced-price lunch (FRL) programs, students with individualized education plans (IEP), and students identified as multilingual learners (ML). Results are reported separately for each of these disaggregated groups in accordance with state and federal accountability requirements.

The overall Framework score is computed using a more complicated scoring process that combines the points earned for results based on each disaggregated group and a small number of additional data points. Importantly, a school is not rated relative to the disaggregated performance of any student groups with fewer than the minimum n-size. For example, if a school has fewer than 16 minority students, the average test scores and SGP of minority students in that school are not used to calculate minority student disaggregated results, although these students' scores would still contribute to the school's overall rating.

Table 1: Scoring Rules for Achievement and Growth

Indicator	Measure/Metric	Points
Achievement	At or above the 85th percentile	8
	At or above the 50th percentile but below the 85th percentile	6
	At or above the 15th percentile but below the 50th percentile	4
	Below the 15th percentile	2
Growth	At or above 65	8
	At or above 50 but below 65	6
	At or above 35 but below 50	4
	Below 35	2

Note. The percentiles used to rate school average achievement are based on the distribution of school average scores rather than student- or district-level distributions.

Although the average test scores and growth metrics contribute indirectly to school and district framework scores via these scoring rules, we analyze properties of the average test scores and SGP rather than framework points. We do this because the statistical properties of these test score metrics are relevant for understanding the eventual Framework ratings and because the average test scores and median SGP are reported publicly alongside the Framework ratings.

Reliability of School Aggregate Test Score Metrics

Reliability refers to the consistency of a set of measurements. In the context of aggregate school test score metrics, reliability is used to assess how consistent the school metrics would be across conceptually equivalent measurement occasions. A reliability coefficient quantifies the proportion of variance in the measurements due to true differences across units that would remain stable across equivalent measurement occasions relative to the proportion of variance due to random factors that would vary across occasions. Quantifying the reliability of observed school (or district) aggregate test scores requires making assumptions about the sources of variability in the observed scores.[5] Statisticians have made a distinction between the "infinite population" model and "finite population" model for characterizing the uncertainty in school test score means (Cronbach et al., 1997). The appropriate model to use depends upon the inferences or generalizations that will be made based on the school mean test scores.

In the *finite population model*, the population of interest is limited to the specific students tested in the school in a specific year. There is assumed to be no sampling error due to the students represented in the mean test score because the tested students constitute the entire population of enrolled students. Under this model, only test score measurement error contributes to variability in a school's mean test score or SGP. In the infinite population model, "an infinite population could be assumed to exist for each school, and the pupils tested could be conceived of as a random sample from the population associated with the school" (Cronbach et al., 1997, p. 391). Under this model, both sampling error due to the particular students tested and test score measurement error contribute to uncertainty in a single school's mean test score or SGP.

Cronbach et al. (1997) argue the infinite population model is more appropriate for supporting the types of inferences made in school accountability systems even when all enrolled students are tested. Cronbach et al. note that, "To conclude on the basis of an assessment that a school is effective as an institution requires the assumption, implicit or explicit, that the positive outcome would appear with a student body other than the present one, drawn from the same population." (p. 393). Hill and DePascale (2003) summarize further arguments in favor of the infinite population model for quantifying uncertainty in average test scores (these arguments apply to aggregate growth metrics as well), and the infinite population model is also consistent with recommendations in the AERA, APA and NCME Standards (2014; Standard 2.17). Seastrom (2017) provides further discussion about adopting the finite versus infinite population models, which are referred to as the "population" and "sampling" models, respectively, concluding that either model can be appropriate for evaluating school accountability metrics depending upon the context.

The infinite population model is an abstract conceptual framework used to connect statistical theory with the ways test scores are used in accountability. By incorporating uncertainty due to student sampling, this framework acknowledges that aggregate test score outcomes may be impacted by idiosyncratic factors specific to the students tested in that year thus providing a way to approximate how much this uncertainty will cause aggregate test score metrics to vary. However, because students are not in fact random samples from well-defined populations, we do not necessarily recommend that standard errors or estimates of uncertainty derived within this framework be used operationally for state accountability systems. We consider the decision about whether to incorporate measures of uncertainty such as those used here into a school accountability system a distinct (though related) matter.

[5] Although the remainder of this section refers to school-level average test scores, the same considerations and calculations can be applied to school-level average growth metrics and to district-level aggregate test score metrics.

Data

We used student-level longitudinal data files provided by CDE to construct a dataset of school and district test score results for the years 2016-17, 2017-18, and 2018-19.[6] Within each year, we included schools and districts with test score data for students in 3rd-8th grade. For average test score analyses, we limited the sample to schools (or districts) that had at least one non-missing test score in both math and ELA. For SGP analyses, we limited the sample to schools (or districts) that had at least one non-missing SGP in both math and ELA. For each school (or district) we calculated the average test score (and number of non-missing test scores) and average SGP (and number of non-missing SGP), separately for math and ELA in each year. We calculated this separately for elementary and middle school grade levels. We do not report results for high schools, although similar analyses could be carried out for high school results. Due to additional business rules applied by CDE in the process of calculating framework results, these school and district test score results may not exactly match those used operationally. However, we expect the results would be similar to those used in the operational calculations.

Table 2: Number of Schools and Districts and Median Sample Sizes

		Achievement				Growth			
		Schools		Districts		Schools		Districts	
Level	**Year**	**N**	**Median**	**N**	**Median**	**N**	**Median**	**N**	**Median**
E	2017	1083	180	175	140	1071	115	173	98
E	2018	1097	180	176	150	1084	116	174	100
E	2019	1105	175	175	150	1098	113	175	103
M	2017	579	197	173	108	575	180	172	96
M	2018	595	194	174	102.5	589	182	172	94
M	2019	596	207	175	105	593	196	174	104

Note. N=number of schools or districts with achievement or growth data; Median=median number of students per school or district.

Table 2 summarizes the number of schools and districts included in the final sample for achievement (average test scores) and growth (SGP) analyses. The table also reports the median number of math test scores or math SGPs observed in each school or district (the number of math and ELA scores is very similar within each school or district). There are about 1,100 elementary schools with non-missing test score data in each year, with median sample sizes of about 180 students for achievement and 115 students for growth. There were slightly fewer than 600 middle schools with non-missing test score data in each year, with median sample sizes of about 200 students for achievement and 180 students for growth. There are data for 172-176 districts at each level across years, with median sample sizes of about 140 students for elementary achievement, 100 students for middle school achievement, and about100 students for growth.

[6] These are the three most recent years of complete data available prior to disruptions to testing caused by the COVID-19 pandemic. Conducting similar analyses with more recent years of data post-pandemic would be a useful extension.

Methods

We use hierarchical linear modeling (HLM; Raudenbush & Bryk, 2002) as a framework to define and estimate the reliability of school aggregate test score metrics. This approach is consistent with the infinite population model described above. We use HLM to estimate the variability in school test score metrics due to true differences between schools relative to random variability. This is a model-based approach to defining the reliability of aggregate school test score metrics. The following model is used:

$$y_{igs} = \gamma_{00} + u_{0s} + e_{igs}$$

$$u_{0s} \sim N(0, \tau_{00})$$

$$e_{igs} \sim N(0, \sigma^2). \qquad \text{Eq. 1}$$

Here, y_{igs} is the test score for student *i* in grade *g* in school *s*. The term γ_{00} is the overall average value across schools, the u_{0s} are normally distributed random school intercepts representing the true mean in each school, and the e_{igs} are normally distributed student-level residuals. The terms τ_{00} and σ^2 represent the variance of the true school means and the student level residuals, respectively. Although the *g* subscripts are redundant with student within year, we include these as a reminder that we are pooling test scores across all relevant tested grades within each school, consistent with the process used for Framework calculations. For example, for elementary schools, we pool test scores across grades 3-5 for a given year, as is done to construct the mean test scores used in the SPF calculations.

The parameter τ_{00} represents the true variance of school means, net of random sampling and measurement error. The reliability of the mean in school *s* is (Raudenbush & Bryk, 2002)

$$Rel(\bar{y}_s) = \frac{\tau_{00}}{\tau_{00} + \frac{\sigma^2}{n_s}}, \qquad \text{Eq. 2}$$

where n_s is the number of tested students in school *s*. Reliability here quantifies how well we can differentiate the true average achievement (net of sampling and measurement error) in school s from true average achievement in other schools. Reliability ranges from 0 to 1, with higher values indicating greater reliability. A reliability of 0.80, for example, suggests that about 80% of the variance in means across schools of the same sample size is due to true differences, while 20% is due to random error.

The reliability depends in part on the true variance of averages across schools. When there are larger true differences in performance across schools (τ_{00} is larger), we can more reliably differentiate performance across schools. Reliability also depends on the term σ^2/n_s which is an estimate of the error for the observed mean in school s, consistent with the infinite population model. Reliability will be higher when test scores vary less within schools or when a school mean is based on more individual test scores. All else equal, schools with more tested students will have more reliable means.

The intraclass correlation coefficient (ICC) quantifies the proportion of variability in test scores (or SGPs) that is between versus within schools, and is defined as

$$ICC = \frac{\tau_{00}}{\tau_{00} + \sigma^2}. \quad \text{Eq. 3}$$

The ICC ranges from 0 to 1. A higher ICC indicates that a greater proportion of variance in test scores is due to true between-school differences. Reliability and the ICC describe related but distinct properties of variability in school mean test scores. The reliability of each school's mean test score depends on the overall ICC and the school sample size; all else equal, school mean test score metrics will be more reliable when the ICC is higher.

Using the test score data, we first calculated the number of schools and districts that would have sufficient sample sizes to report data under different minimum n-size thresholds. Next, we estimated the HLM model in Equation 1 with student scale scores as the outcome (y_{igs}) for each subject (math or ELA), year (2017-2019), and grade level (elementary or middle school) separately. We did this first aggregating at the school level and then aggregating at the district level. We used the parameter estimates to calculate the ICC and reliability values of interest. We used the same models and methods to estimate the reliability of school mean SGP by replacing y_{igs} in Equation 1 with SGP_{igs}, the SGP for each student.[7]

These analyses are first conducted based on all students in each school or district. We then repeat the analyses based on the four disaggregated groups listed above that are used in Framework calculations. The HLM models were estimated using a sample limited to schools with at least five non-missing test scores or SGPs. All analysis was carried out using the R statistical software (R Core Team, 2023). HLM models were estimated using the "lmer" package (Bates et al., 2015).

[7] These models are estimated using maximum likelihood estimation (ML) assuming that the distribution of true school means and the student-level within-school residuals are each normally distributed. While this assumption is a reasonable approximation for scale scores, the within-school distribution of SGP is not well approximated by a normal distribution. We are investigating the ways this might impact results and alternative strategies to address this concern.

Results

Sample Size Impacts on Available Data

Table 3 reports the number and percent of schools and districts in 2019 that had sufficient data to report results with the current minimum n-size threshold of 16 (for achievement) or 20 (for growth), along with the number and percent of schools that would have sufficient data to report results if both minimum n-size thresholds were reduced to n=10 students.[8] The first row of Table 3, for example, indicates that of the 1,105 elementary schools with test score data in 2019, 1,068 (97.6%) had at least 16 test scores in each of ELA and math and thus would have data publicly reported. If the minimum n-size threshold were reduced to N=10 students, there would be 1,088 (98.5%) elementary schools with sufficient data for reporting. Reducing the minimum n-size to N=10 would lead to reporting data for an additional 20 schools (1.8% of all schools). At the district level, reducing the minimum n-size from 16 to 10 would lead to reporting achievement data for an additional 9 districts (5.1% of all districts). The results described here were similar for 2017 and 2018; detailed tables are provided in the Appendix.

For achievement data, where the current minimum n-size is 16, over 90% of schools and districts at both the elementary and middle school levels had sufficient sample sizes for reporting results. Lowering the minimum n-size to 10 would lead to reporting data for about 2-3% more schools and about 4-5% more districts for overall results. For disaggregated groups, lowering the minimum n-size results in greater increases in the number for schools and districts with reportable achievement data, and the increases vary across groups and grade levels. In contrast to the increases in reporting for all students, there would be larger proportional increases in the amount of school data reported for disaggregated groups relative to districts. The largest increases would be for achievement data of students identified as multilingual learners and students with IEPs.

For growth data, reducing the minimum n-size from 20 to 10 would lead to larger increases in the number and percent of schools and districts with reportable data, particularly for disaggregated groups. Data for all students would be reported for an additional 2-5% of schools and 6-7% of districts. Increases in the number of schools and districts with reportable disaggregated group data would be even larger, ranging from about 6% to as many as 45% of additional schools. For example, an additional 498 elementary schools (45.4% of all schools) would have reportable growth data for students with IEPs. However, this is a much larger increase than would occur for other disaggregated groups at both the school and district levels, where increases are 10-20%.

Overall, reducing the minimum n-size to 10 students for achievement and growth would lead to modest increases in the number of schools and districts with reportable data for overall student performance. The change would lead to more substantial increases in the number of schools and districts with reportable data for disaggregated groups, particularly growth data for students with an IEP in elementary schools.

[8] The 1241 Task Force report includes similar calculations in Appendix A. Although the results in the Task Force report are aggregated differently, the patterns are similar.

Table 3: Number of Schools and Districts Meeting Different Minimum Sample Size Requirements in 2019, by Year and Level

		Schools			Districts		
		Achievement					
Level	Group	N>=10	N>=16	Change	N>=10	N>=16	Change
E	All	1088 (98.5)	1068 (96.7)	1.8	171 (97.7)	162 (92.6)	5.1
E	Minority	1004 (90.9)	966 (87.4)	3.4	130 (74.3)	114 (65.1)	9.1
E	IEP	888 (80.4)	721 (65.2)	15.1	110 (62.9)	93 (53.1)	9.7
E	FRL	1000 (90.5)	923 (83.5)	7.0	161 (92.0)	147 (84.0)	8.0
E	ML	667 (60.4)	539 (48.8)	11.6	83 (47.4)	77 (44.0)	3.4
M	All	571 (95.8)	556 (93.3)	2.5	167 (95.4)	159 (90.9)	4.6
M	Minority	503 (84.4)	463 (77.7)	6.7	130 (74.3)	112 (64.0)	10.3
M	IEP	397 (66.6)	318 (53.4)	13.3	104 (59.4)	82 (46.9)	12.6
M	FRL	501 (84.1)	461 (77.3)	6.7	154 (88.0)	141 (80.6)	7.4
M	ML	352 (59.1)	307 (51.5)	7.6	82 (46.9)	69 (39.4)	7.4
		Growth					
Level	Group	N>=10	N>=20	Change	N>=10	N>=20	Change
E	All	1063 (96.8)	1037 (94.4)	2.4	162 (92.6)	150 (85.7)	6.9
E	Minority	959 (87.3)	867 (79.0)	8.4	115 (65.7)	104 (59.4)	6.3
E	IEP	679 (61.8)	181 (16.5)	45.4	94 (53.7)	68 (38.9)	14.9
E	FRL	919 (83.7)	763 (69.5)	14.2	146 (83.4)	121 (69.1)	14.3
E	ML	552 (50.3)	349 (31.8)	18.5	78 (44.6)	66 (37.7)	6.9
M	All	568 (95.8)	538 (90.7)	5.1	165 (94.8)	154 (88.5)	6.3
M	Minority	492 (83.0)	429 (72.3)	10.6	126 (72.4)	102 (58.6)	13.8
M	IEP	378 (63.7)	269 (45.4)	18.4	99 (56.9)	67 (38.5)	18.4
M	FRL	489 (82.5)	428 (72.2)	10.3	151 (86.8)	128 (73.6)	13.2
M	ML	346 (58.3)	281 (47.4)	11.0	80 (46.0)	65 (37.4)	8.6

Note. Data are based on spring 2019 test results. Values in the "N>=" columns report the count of schools or districts with at least N non-missing scores or SGP in math and in ELA (values in parentheses report what percent of all schools or districts with data these represent). Values in the "Change" columns report the increase in percent of schools or districts with sufficient data for public reporting. FRL=free or reduced-price lunch; ML=multilingual learners; IEP=individualized education plan.

ICC Estimates

Table 4 reports ICCs estimates based on 2019 data across schools and districts, separately by subject and disaggregated groups for achievement (scale scores) and growth (SGP). As a reminder, the ICC quantifies the proportion of variance in test scores (or SGP) that is between schools rather than within. When there is less variance between schools (i.e., a lower ICC), it is more difficult to reliably differentiate performance across schools based on test scores or SGP.

Among schools, the average ICC across subjects, groups, and grade levels is 0.15 for achievement and 0.06 for growth. Hence, approximately 15% of the variance in test scores is due to differences between schools, while only about 6% of the variance in SGPs is due to differences between schools. For districts, the average ICCs are about 0.06 for achievement and 0.05 for growth. The magnitude of achievement ICCs in Table 4 are consistent with prior literature reporting estimates of between-district and between-school ICCs in test scores (e.g., Fahle & Reardon, 2018; Hedges & Hedberg, 2007, 2014). We are not aware of prior studies reporting ICC values for SGP.

Comparing across subjects, achievement ICCs at the school level are consistently higher in math than ELA, while growth ICCs are similar across subjects in middle school but consistently higher in math among elementary schools. At the district level, achievement ICCs are similar across subjects, while growth ICCs tend to be higher in math than ELA. Disaggregated group ICCs follow similar patterns as those for results based on all students but tend to be smaller than the ICCs based on all students.

Table 4: Estimated ICCs Across Schools and Districts, by Year and Level

		Schools			Districts		
		Achievement					
Level	Group	N	ELA	Math	N	ELA	Math
E	All	1097	0.17	0.19	174	0.08	0.08
E	Minority	1040	0.15	0.18	147	0.06	0.07
E	IEP	1003	0.12	0.13	136	0.06	0.05
E	FRL	1053	0.09	0.11	172	0.05	0.06
E	ML	837	0.15	0.19	100	0.05	0.06
M	All	587	0.17	0.20	172	0.08	0.09
M	Minority	540	0.17	0.20	149	0.07	0.07
M	IEP	481	0.10	0.10	131	0.05	0.06
M	FRL	545	0.11	0.12	167	0.06	0.06
M	ML	406	0.16	0.19	99	0.05	0.05
		Growth					
Level	Group	N	ELA	Math	N	ELA	Math
E	All	1082	0.06	0.09	171	0.04	0.08
E	Minority	1013	0.05	0.08	137	0.03	0.07
E	IEP	924	0.05	0.05	118	0.01	0.05
E	FRL	1012	0.05	0.08	162	0.03	0.06
E	ML	744	0.05	0.08	93	0.01	0.06
M	All	582	0.07	0.07	171	0.05	0.09
M	Minority	536	0.07	0.07	147	0.03	0.07
M	IEP	464	0.04	0.02	123	0.01	0.06
M	FRL	539	0.07	0.06	162	0.04	0.06
M	ML	399	0.07	0.07	97	0.02	0.05

Note. Values based on spring 2019 test score data. The sample was limited to units with at least 5 non-missing scale scores or SGPs when calculating ICCs. The values in "N" indicate the total number of schools or districts included in the HLM analysis to estimate ICCs. FRL=free or reduced-price lunch; ML=multilingual learners; IEP=individualized education plan.

The primary takeaway from these results is that larger sample sizes will be needed to achieve the same level of reliability for growth metrics (average SGP) relative to achievement metrics (average scale score) and for disaggregated groups relative to all students, because the ICCs for growth metrics and for disaggregated groups are lower. The growth ICC should be interpreted somewhat cautiously given the caveat noted earlier that the models used to estimate ICCs assume scores or SGP are normally distributed. In addition to the modeling assumptions, there are two primary reasons that growth ICCs might be lower than achievement ICCs. First, there could be smaller true differences in student growth across schools than there are in student achievement. Second, student growth measures tend to have larger measurement error than status measures. Measurement error will tend to inflate the within-school and within-district variance, thus reducing the ICC. These explanations are not mutually exclusive.

Reliability Estimates

Table 5 reports the estimated reliability of group averages under different combinations of ICC and sample size. As a reminder, the reliability coefficient quantifies the proportion of variance in average school or district metrics that is due to consistent true differences between schools or districts rather than random error. The estimated reliability for schools of a particular sample size can be calculated based on the ICC.[9] The ICC values in Table 5 were selected to represent the range of values observed across different groups, units, and subjects. For example, with an ICC of 0.20 (the highest school-level achievement ICC, observed for middle school math scores) and 10 tested students, the reliability of the school mean is about 0.71; the expected reliability increases to 0.80 when there are 16 tested students. These reliability estimates would be 0.34 and 0.46 for an ICC of 0.05, which is slightly below the average achievement ICC observed at the district level. Figure 1 shows these results graphically, and illustrates that when ICCs are lower, larger sample sizes are needed to achieve the same level of reliability. The figure also emphasizes that the increase in reliability for each additional test score (or SGP) is greatest when sample sizes are below 20.

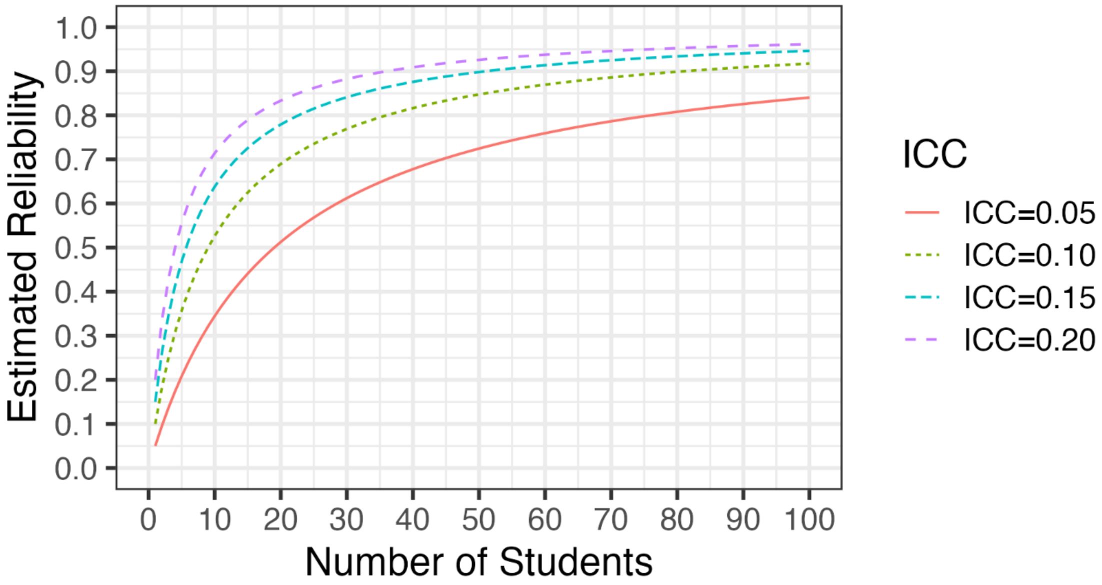
Table 5: Implied Reliability by Sample Size and ICC Value

	Reliability when...			
N	**ICC=0.05**	**ICC=0.10**	**ICC=0.15**	**ICC=0.20**
5	0.21	0.36	0.47	0.56
10	0.34	0.53	0.64	0.71
15	0.44	0.62	0.73	0.79
16	0.46	0.64	0.74	0.80
20	0.51	0.69	0.78	0.83
36	0.65	0.80	0.86	0.90
50	0.72	0.85	0.90	0.93
100	0.84	0.92	0.95	0.96
150	0.89	0.94	0.96	0.97
200	0.91	0.96	0.97	0.98

[9] Reliability can be calculated from the ICC and sample size as:

$$reliability = \frac{ICC}{ICC + \frac{(1 - ICC)}{N}}$$

Figure 1: Estimated Reliability by Sample Size and ICC Value



The average reliability across all schools and districts will be higher than the values in Table 5, which are reported for specific, mostly small sample sizes. The median school has test scores for about 175-200 students and SGPs for about 115-200 students, while the median district has test scores for 100-150 students and SGPs for about 100 students. With an ICC of 0.05 (the lower end of values observed in Table 4) and a sample size of 100, the reliability of group means is about 0.84; estimated reliability would be greater than 0.90 for ICCs of 0.10 or higher. With a sample size of 150 students, estimated reliability would range from 0.89 (ICC=0.05) to 0.97 (ICC=0.20) and with a sample size of 200 students, estimated reliability would range from 0.91 to 0.98.

The relationship between sample size, ICC, and reliability highlights the tradeoff with reducing the minimum n-size. For example, while reducing the minimum n-size to N=10 would require many more elementary schools to report growth data for students with IEPs, the ICC of average SGP among elementary schools for students with IEPs was 0.05. With an ICC of 0.05 and N=10, the reliability is 0.34. In other words, only about 34% of the variation in average SGP across schools is due to true differences in student performance when inferences are based on only 10 observed SGP.

Discussion

Setting a minimum n-size threshold for accountability reporting involves making tradeoffs between transparency and reliability. A clearly articulated framework for defining and quantifying reliability should be used to inform setting a minimum n-size, whether it is the approach described here or an alternative. Lower minimum n-size thresholds will enable reporting data for more units (particularly for disaggregated groups within schools and districts), but data reported for small groups or units may be less reliable. Although it is straightforward to determine how many additional schools or districts will have sufficient data under different n-size thresholds, quantifying the impact on reliability requires defining reliability.

The increase in the percentage of schools or districts with reportable data when reducing the minimum n-size to 10 students for achievement and growth was modest for overall unit results. We would expect increases of about 2-7 percentage points in the percent of schools and districts reporting data for all students. The increases were larger for disaggregated group results, where sample sizes tend to be smaller, although the increases varied considerably across groups, levels and metrics. Determining whether these increases in reporting would be worth the reduction in reliability requires consideration about how these results would be used to improve educational opportunities for students enrolled in the schools and districts that would have additional reportable data.

The relative variability among average scale scores, quantified by the ICC, was consistently larger than the relative variability among average SGPs. As a result, the reliability of average test scores was higher than the reliability of average SGP for equivalent sample sizes. This supports the current policy to have a higher minimum n-size for growth metrics (n=20) than for achievement metrics (n=16). The school level ICCs were greater than the district level ICC; thus, for the same sample size, aggregate school metrics were more reliable than aggregate district metrics. This is consistent with prior research documenting higher ICCs at the school level than the district level, and may arise because, relative to schools, districts on average enroll more heterogeneous student populations from larger geographic areas.

Because the reliability statistics we calculate are sample size dependent, the reliabilities reported for a particular minimum n-size (e.g., n=16) should be considered a lower bound. Most units will have larger sample sizes than the minimum and thus have more reliable estimates. These calculations could inform setting a minimum n-size by determining what the lowest anticipated reliability would be for various minimum n-sizes. In addition, because reliability depends on the distribution of outcomes across units, the reliability at a fixed sample size could vary from year to year, for schools versus districts, and for different disaggregated groups. Norm-referenced thresholds for assigning SPF and DPF points are set relative to a baseline dataset; a similar dataset could be used to determine the minimum n-sizes necessary to achieve a desired minimum level of reliability.

We also note some limitations and potential extensions to this work. First, having a sufficient sample size for reliable aggregate metrics does not guarantee that the metrics will be valid indicators for the intended inferences. While a high degree of reliability may be considered a necessary condition to ensuring the indicators are valid, it is not a sufficient condition. Second, the ICC estimates were based on samples limited to units with at least five non-missing test scores or SGPs. Depending upon the inferences one is making it may be more appropriate to include all units or estimate ICCs limiting only to units with sufficient data for reporting. Finally, the models used to estimate ICCs assume that student-level scores are normally distributed and are based on comparisons of mean SGP not median SGP (the median SGP is currently used in SPF calculations). Further work could estimate the ICC and reliability using alternative methods to gauge the reliability of median SGP and account for non-normality of the SGP.

There are two extensions to these analyses that could be pursued. First, researchers have recently investigated methods for “stabilizing” school or district aggregate test score metrics (Castellano et al., 2023; Forrow et al., 2023; Lockwood et al., 2022; Rosendahl et al., 2024). These approaches use auxiliary information (often data from each unit’s prior historical performance) to adjust aggregate test score metrics, thereby reducing random error and increasing reliability and accuracy. These approaches carry different strengths and limitations that could be considered for the context of Colorado SPF ratings. Second, although improving reliability of aggregate metrics is useful, future analyses could also evaluate the precision of unit aggregate test score metrics based on the standard error of measurement (SEM). The aggregate test score metrics in Colorado are compared to fixed cut scores to assign SPF ratings based on which score band the aggregate value falls within (see Table 1). In these contexts, it is helpful to evaluate the magnitude of error in both absolute terms (based on the SEM) and in relative terms (based on the reliability). If the scoring bands are wide relative to the variability of true group means, for example, it would be possible for group means to be unreliable but still be accurately located within one of the scoring bands. Conversely, if the scoring bands are narrow relative to the variability in group means, then it would be possible to have reliable group means but be unable to accurately locate group means in the correct scoring band.

Appendix

Table A1: Number of Schools and Districts Meeting Different Minimum Sample Size Requirements in 2017, by Group and Level

		Schools			Districts		
		Achievement					
Level	**Group**	**N>=10**	**N>=16**	**Change**	**N>=10**	**N>=16**	**Change**
E	All	1057 (97.6)	1040 (96.0)	1.57	165 (94.3)	157 (89.7)	4.57
E	Minority	978 (90.3)	945 (87.3)	3.05	128 (73.1)	118 (67.4)	5.71
E	IEP	848 (78.3)	667 (61.6)	16.71	104 (59.4)	87 (49.7)	9.71
E	FRL	962 (88.8)	906 (83.7)	5.17	155 (88.6)	145 (82.9)	5.71
E	ML	687 (63.4)	567 (52.4)	11.08	88 (50.3)	79 (45.1)	5.14
M	All	546 (94.3)	529 (91.4)	2.94	162 (93.6)	155 (89.6)	4.05
M	Minority	482 (83.2)	441 (76.2)	7.08	127 (73.4)	113 (65.3)	8.09
M	IEP	361 (62.3)	296 (51.1)	11.23	93 (53.8)	74 (42.8)	10.98
M	FRL	477 (82.4)	429 (74.1)	8.29	147 (85.0)	135 (78.0)	6.94
M	ML	351 (60.6)	314 (54.2)	6.39	84 (48.6)	73 (42.2)	6.36
		Growth					
Level	**Group**	**N>=10**	**N>=20**	**Change**	**N>=10**	**N>=20**	**Change**
E	All	1029 (96.1)	1005 (93.8)	2.24	155 (89.6)	147 (85.0)	4.62
E	Minority	934 (87.2)	830 (77.5)	9.71	118 (68.2)	98 (56.6)	11.56
E	IEP	615 (57.4)	146 (13.6)	43.79	85 (49.1)	59 (34.1)	15.03
E	FRL	902 (84.2)	754 (70.4)	13.82	146 (84.4)	117 (67.6)	16.76
E	ML	585 (54.6)	371 (34.6)	19.98	81 (46.8)	68 (39.3)	7.51
M	All	539 (93.7)	510 (88.7)	5.04	160 (93.0)	149 (86.6)	6.40
M	Minority	469 (81.6)	413 (71.8)	9.74	123 (71.5)	102 (59.3)	12.21
M	IEP	343 (59.7)	237 (41.2)	18.43	88 (51.2)	61 (35.5)	15.70
M	FRL	466 (81.0)	394 (68.5)	12.52	145 (84.3)	121 (70.3)	13.95
M	ML	346 (60.2)	292 (50.8)	9.39	83 (48.3)	71 (41.3)	6.98

Note. Data are based on spring 2017 test results. Values in the "N>=" columns report the count of schools or districts with at least N non-missing scores or SGP in math and in ELA (values in parentheses report what percent of all schools or districts with data these represent). Values in the "Change" columns report the increase in percent of schools or districts with sufficient data for public reporting. FRL=free or reduced-price lunch; ML=multilingual learners; IEP=individualized education plan.

Table A2: Number of Schools and Districts Meeting Different Minimum Sample Size Requirements in 2018, by Group and Level

		Schools			Districts		
		Achievement					
Level	**Group**	**N>=10**	**N>=16**	**Change**	**N>=10**	**N>=16**	**Change**
E	All	1074 (97.9)	1061 (96.7)	1.19	169 (96.0)	163 (92.6)	3.41
E	Minority	996 (90.8)	955 (87.1)	3.74	131 (74.4)	116 (65.9)	8.52
E	IEP	866 (78.9)	705 (64.3)	14.68	108 (61.4)	92 (52.3)	9.09
E	FRL	992 (90.4)	928 (84.6)	5.83	161 (91.5)	147 (83.5)	7.95
E	ML	695 (63.4)	558 (50.9)	12.49	86 (48.9)	77 (43.8)	5.11
M	All	562 (94.5)	546 (91.8)	2.69	163 (93.7)	157 (90.2)	3.45
M	Minority	495 (83.2)	459 (77.1)	6.05	122 (70.1)	113 (64.9)	5.17
M	IEP	388 (65.2)	320 (53.8)	11.43	96 (55.2)	83 (47.7)	7.47
M	FRL	502 (84.4)	458 (77.0)	7.39	155 (89.1)	139 (79.9)	9.2
M	ML	355 (59.7)	312 (52.4)	7.23	84 (48.3)	74 (42.5)	5.75
		Growth					
Level	**Group**	**N>=10**	**N>=20**	**Change**	**N>=10**	**N>=20**	**Change**
E	All	1050 (96.9)	1017 (93.8)	3.04	159(91.4)	145 (83.3)	8.05
E	Minority	954 (88.0)	854 (78.8)	9.23	117(67.2)	102 (58.6)	8.62
E	IEP	658 (60.7)	173 (16.0)	44.74	92(52.9)	65 (37.4)	15.52
E	FRL	918 (84.7)	763 (70.4)	14.3	144(82.8)	118 (67.8)	14.94
E	ML	574 (53.0)	371 (34.2)	18.73	80(46)	67 (38.5)	7.47
M	All	553 (93.9)	528 (89.6)	4.24	158(91.9)	148 (86.0)	5.81
M	Minority	482 (81.8)	429 (72.8)	9.00	121(70.3)	105 (61.0)	9.30
M	IEP	361 (61.3)	255 (43.3)	18.00	86(50)	65 (37.8)	12.21
M	FRL	485 (82.3)	423 (71.8)	10.53	148(86)	126 (73.3)	12.79
M	ML	346 (58.7)	279 (47.4)	11.38	82(47.7)	64 (37.2)	10.47

Note. Data are based on spring 2018 test results. Values in the “N>=” columns report the count of schools or districts with at least N non-missing scores or SGP in math and in ELA (values in parentheses report what percent of all schools or districts with data these represent). Values in the “Change” columns report the increase in percent of schools or districts with sufficient data for public reporting. FRL=free or reduced-price lunch; ML=multilingual learners; IEP=individualized education plan.

Table A3: Estimated ICCs Across Schools and Districts, by Group and Level (2017)

		Schools			Districts		
		Achievement					
Level	Group	N>=10	N>=16	Change	N>=10	N>=16	Change
E	All	1072	0.18	0.2	172	0.09	0.1
E	Minority	1019	0.16	0.19	151	0.08	0.09
E	IEP	965	0.13	0.14	133	0.08	0.07
E	FRL	1021	0.1	0.11	167	0.08	0.09
E	ML	837	0.16	0.2	109	0.07	0.07
M	All	566	0.19	0.21	169	0.1	0.09
M	Minority	514	0.18	0.2	141	0.08	0.07
M	IEP	453	0.11	0.13	120	0.05	0.06
M	FRL	524	0.13	0.12	165	0.07	0.06
M	ML	405	0.18	0.2	100	0.06	0.07
		Growth					
Level	Group	N>=10	N>=20	Change	N>=10	N>=20	Change
E	All	1054	0.08	0.09	167	0.08	0.1
E	Minority	984	0.06	0.09	138	0.05	0.09
E	IEP	874	0.05	0.06	115	0.03	0.07
E	FRL	981	0.06	0.08	161	0.06	0.09
E	ML	755	0.06	0.1	99	0.04	0.07
M	All	561	0.1	0.08	166	0.09	0.09
M	Minority	505	0.1	0.07	138	0.06	0.07
M	IEP	440	0.06	0.03	117	0.02	0.06
M	FRL	511	0.1	0.07	162	0.07	0.06
M	ML	393	0.1	0.07	95	0.05	0.07

Note. Values based on spring 2017 test score data. The sample was limited to units with at least 5 non-missing scale scores or SGPs when calculating ICCs. The values in "N" indicate the total number of schools or districts included in the HLM analysis to estimate ICCs. FRL=free or reduced-price lunch; ML=multilingual learners; IEP=individualized education plan.

Table A4: Estimated ICCs Across Schools and Districts, by Group and Level (2018)

		Schools			Districts		
		Achievement					
Level	Group	N>=10	N>=16	Change	N>=10	N>=16	Change
E	All	1087	0.17	0.19	173	0.07	0.09
E	Minority	1039	0.16	0.18	154	0.07	0.08
E	IEP	992	0.12	0.13	135	0.06	0.07
E	FRL	1041	0.09	0.11	170	0.05	0.08
E	ML	842	0.14	0.19	105	0.05	0.07
M	All	582	0.18	0.2	170	0.09	0.09
M	Minority	530	0.17	0.19	145	0.07	0.07
M	IEP	470	0.1	0.12	124	0.05	0.06
M	FRL	536	0.12	0.12	163	0.06	0.05
M	ML	417	0.15	0.18	101	0.04	0.05
		Growth					
Level	Group	N>=10	N>=20	Change	N>=10	N>=20	Change
E	All	1073	0.06	0.1	170	0.04	0.09
E	Minority	1002	0.06	0.09	136	0.03	0.08
E	IEP	894	0.05	0.05	111	0.02	0.07
E	FRL	998	0.05	0.08	162	0.03	0.08
E	ML	767	0.06	0.09	94	0.02	0.07
M	All	572	0.08	0.06	167	0.05	0.09
M	Minority	523	0.08	0.06	141	0.03	0.07
M	IEP	448	0.05	0.03	117	0.02	0.06
M	FRL	528	0.08	0.06	160	0.04	0.05
M	ML	406	0.07	0.07	97	0.02	0.05

Note. Values based on spring 2018 test score data. The sample was limited to units with at least 5 non-missing scale scores or SGPs when calculating ICCs. The values in “N” indicate the total number of schools or districts included in the HLM analysis to estimate ICCs. FRL=free or reduced-price lunch; ML=multilingual learners; IEP=individualized education plan.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Alliance for Excellent Education. (2018, December 7). *N-size in ESSA state plans.* https://all4ed.org/publication/n-size-in-essa-state-plans/

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1). https://doi.org/10.18637/jss.v067.i01

Betebenner, D. W. (2009). Norm-and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42–51. https://doi.org/10.1111/j.1745-3992.2009.00161.x

Castellano, K. E., & McCaffrey, D. F. (2017). The accuracy of aggregate student growth percentiles as indicators of educator performance. *Educational Measurement: Issues and Practice, 36*(1), 14–27. https://doi.org/10.1111/emip.12144

Castellano, K. E., McCaffrey, D. F., & Lockwood, J. R. (2023). An exploration of an improved aggregate student growth measure using data from two states. *Journal of Educational Measurement, 60*(2), 173–201. https://doi.org/10.1111/jedm.12354

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability analysis for performance assessments of student achievement or school effectiveness. *Educational and Psychological Measurement, 57*(3), 373–399. https://doi.org/10.1177/0013164497057003001

Every Student Succeeds Act, Pub. L. No. 114–95 (2015).

Fahle, E. M., & Reardon, S. F. (2018). How much do test scores vary among school districts? New estimates using population data, 2009–2015. *Educational Researcher, 47*(4), 221–234. https://doi.org/10.3102/0013189X18759524

Forrow, L., Starling, J., & Gill, B. (2023). *Stabilizing subgroup proficiency results to improve the identification of low-performing schools* (No. REL 2023-001). US Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/rel/Products/Publication/106926

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60–87. https://doi.org/10.3102/0162373707299706

Hedges, L. V., & Hedberg, E. C. (2014). Intraclass correlations and covariate outcome correlations for planning two- and three-level cluster-randomized experiments in education. *Evaluation Review, 37*(6), 445–489. https://doi.org/10.1177/0193841X14529126

Hill, R. K., & DePascale, C. A. (2003). Reliability of No Child Left Behind accountability designs. *Educational Measurement: Issues and Practice, 22*(3), 12–20. https://doi.org/10.1111/j.1745-3992.2003.tb00133.x

Lockwood, J. R., Castellano, K. E., & McCaffrey, D. F. (2022). Improving accuracy and stability of aggregate student growth measures using empirical best linear prediction. *Journal of Educational and Behavioral Statistics, 47*(5), 544–575. https://doi.org/10.3102/10769986221101624

R Core Team. (2023). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Sage Publications, Inc.

Rosendahl, M., Gill, B., & Starling, J. E. (2024). *Stabilizing school performance indicators in New Jersey to reduce the effect of random error* (No. REL 2025-009). 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/108130

Seastrom, M. (2017). *Best practices for determining subgroup size in accountability systems while protecting personally identifiable student information* (No. IES 2017-147). US Department of Education, Institute of Education Sciences. https://nces.ed.gov/pubs2017/2017147.pdf