H.B. 25-1278 Small Systems Accountability Study Legislative Report

Submitted to:

Office of the Governor

Colorado State Board of Education

House and Senate Education Committees of the Colorado General Assembly

By:

Anne Laesecke, Director, Accountability Analytics; Marie Huchton, Supervisor, Accountability Analytics; Daniel Mangan, Ph.D., Research Associate, Center for Assessment, Design, Research & Evaluation (CADRE)

November 2026

Accountability Analytics Unit
Assessment and Accountability Analytics Division, Colorado Department of Education
201 E. Colfax Ave., Denver, CO 80203
laesecke_a@cde.state.co.us

Table of Contents

Executive Summary 3

Key Findings 3

Stakeholder Feedback 3

Recommendations 3

Introduction & Background 4

Research Questions and Methods 4

Results 5

Lowering Student Count Thresholds Study Findings 6

Measure Stabilization Study Findings 6

Stakeholder feedback 8

TAP feedback 8

AWG feedback 8

Rural Education Council Feedback 9

Recommendations 9

Conclusion 10

References 11

Acronyms

Full name	Acronym
Accountability Work Group	AWG
Center for Assessment, Design, Research and Evaluation	CADRE
Colorado Department of Education	CDE
District Performance Framework	DPF
Empirical Best Linear Predictor	EBLP
Intraclass Correlation	ICC
Personally Identifiable Information	PII
School Performance Framework	SPF
Student Growth Percentile	SGP
Technical Advisory Panel	TAP

Executive Summary

This report meets requirements laid out in H.B. 25-1278 for the Colorado Department of Education (CDE) to conduct two studies into accountability calculation and reporting approaches for small systems. To do this, CDE partnered with the Center for Assessment, Design, Research and Evaluation (CADRE) at the University of Colorado Boulder to conduct these two studies. This report reviews findings from these studies, encapsulates feedback from key stakeholders, and makes recommendations for accountability calculations based on findings.

Key Findings

Sample Size (N-Size) Dynamics	Lowering the minimum N-size threshold for accountability calculations modestly increases overall school data reporting and significantly improves visibility for disaggregated student subgroups. However, smaller sample sizes decrease statistical reliability and increase year-to-year score volatility, particularly for growth metrics.
Statistical Stabilization (EBLP)	Sites with small populations of students experience outsized swings in ratings with changes in student inclusions. Applying Empirical Best Linear Prediction (EBLP) stabilization reduces score volatility and improves underlying test score accuracy for small systems. However, EBLP does not necessarily improve the stability of final performance framework ratings.

Stakeholder Feedback

The **Technical Advisory Panel** recommended maintaining Colorado's current n-size thresholds and calculation methods without statistical stabilization, concluding that the added complexity and potential volatility of changes outweigh the benefits and threaten data reliability, transparency, and consistency.	The **Accountability Work Group** urged CDE to pivot from framework calculation changes toward practice-based support for small and rural districts through clear communication tools, flexible data aggregation models, and the elevation of best practices in small sites for underserved student groups.	Members of the **Rural Education Council** voiced sentiments against lowering n-size thresholds or altering calculation methods in high-stakes accountability frameworks, arguing that the resulting data unreliability could unfairly trigger severe consequences.

Recommendations

- **Do not lower minimum reporting thresholds.** Setting a minimum n-size threshold for accountability reporting involves tradeoffs between transparency and reliability, which stakeholders felt did not warrant a lower reporting threshold.
- **Do not adopt a stabilization methodology.** Adopting Empirical Best Linear Predictor (EBLP), an approach that produces a stabilized test score metric by combining a school's current-year average with additional information in state performance frameworks, is not recommended due to significant trade-offs with transparency, operational complexity and administrative costs.
- **Support small systems in other facets of the accountability system.** CDE should consider addressing other components of the state accountability system to effectively support small systems. This approach should emphasize public reporting for small systems, highlight sites that successfully support historically underserved student groups, and include more customized guidance and technical assistance related to the operational realities of small systems.

Introduction & Background

Schools and districts with small student populations (i.e., small systems) play a vital role in Colorado's public education ecosystem. As highlighted by the H.B. 23-1241 Task Force, accountability measures pose challenges for small systems, including statistical volatility in year-over-year performance ratings and trade-offs surrounding minimum reporting thresholds (N-sizes) that risk masking equity gaps between disaggregated groups. Because these issues are compounded by administrative resource constraints, the task force felt these issues warrant further study to ensure the state's performance frameworks evaluate small educational communities fairly and accurately. Improving accountability systems for small schools, districts, and student groups requires balancing greater transparency while maintaining reliable and stable performance measures. Thus, it is important to note that the 1241 Taskforce recommendations both to lower reporting thresholds and address volatility in test scores for small groups are in tension with one another.

This report is one piece of the Colorado Department of Education's (CDE) work to support small systems. Other efforts include creating multi-year frameworks to aggregate data over time and reach reporting thresholds, which allows CDE to generate reports for as many sites as possible. CDE also allows districts to select final district or school ratings when sites do not have tested grades or have small populations of tested students that do not meet reporting thresholds. To support data access, CDE has created the Secure Data Explorer Tool for districts to analyze state data at the individual student level, even in sites that do not meet reporting thresholds. Finally, CDE has developed improvement planning guidance on how small systems can engage in continuous improvement without reporting Personally Identifiable Information (PII).

Key Terms

- **Small system:** a general term used to capture schools and districts with small student populations. While CDE does not have an official definition of a small system, this usually refers to schools or districts with fewer than 150 students.
- **Minimum N:** The minimum number of students required for public reporting. Currently, the minimum N for achievement is 16 and the minimum N for growth is 20.
- **Reliability:** In statistics, reliability helps us understand how consistent a school's results are over time. A single year's score reflects both the school's performance and the circumstances of that year — which students were tested, how they felt that day, random variation in the questions asked. Reliability asks: would we see the same thing if we looked again at another time?
- **Score volatility**: Year-to-year changes in reported school or district test score metrics that can occur even when underlying performance is relatively stable.
- **Measure stabilization:** A range of statistical techniques that can smooth random fluctuations in test scores due to changes in cohort, test items, or other factors outside of a school's control.

H.B.25-1278 directs CDE to conduct studies and make recommendations on lowering student count thresholds for accountability calculations and reporting and addressing inherent volatility of test score measurements for local education providers with small student populations. This report meets these requirements and outlines research questions and methods used in the corresponding analyses. The report reviews findings from these studies, summarizes recommendations for accountability calculations based on study findings, and encapsulates feedback from various stakeholder groups including the Technical Advisory Panel (TAP), Accountability Work Group (AWG), and Rural Education Council.

Research Questions and Methods

CDE partnered with the Center for Assessment, Design, Research and Evaluation (CADRE) at the University of Colorado Boulder to conduct studies on lowering student reporting thresholds and addressing score volatility

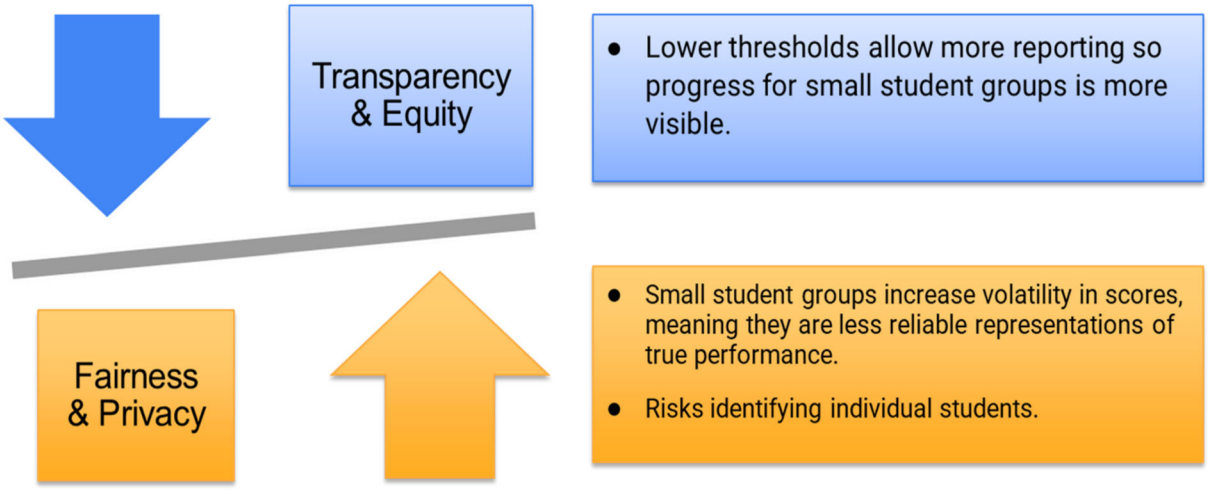
through a stabilization approach called Empirical Best Linear Predictor (EBLP). The EBLP approach produces a stabilized test score metric by combining a school's current-year average with additional information (primarily the school's own results from prior years) in a way that depends on how precisely the current-year value is estimated. EBLP was selected among a variety of stabilization techniques as the most appropriate option for CDE to explore. Another method considered was Bayesian stabilization; however, this was not included as it uses aggregate rather than student-level data; adjusts the state average rather than school-specific data; and had not been operationally adopted by any states. Shrinkage within value-added models was also considered; however, as Colorado does not use a value-added growth model, it could not be directly applied to our data.

Both studies used CDE accountability data and followed CDE business rules for accountability calculations to produce school-level indicators (mean test scores or mean and median SGPs) and performance ratings. These rules establish minimum thresholds of 16 students for achievement indicators and 20 students for growth indicators. The studies then explored impacts on reporting coverage and reliability using a minimum threshold of 10 students for both achievement and growth indicators. More detailed explanations of studies' methodologies and analytical approaches can be found in the technical reports.

Results

To ensure Colorado's SPFs and DPFs accurately and equitably reflect school quality, analyses evaluated two distinct policy mechanisms: lowering minimum student count thresholds to expand reporting coverage and applying EBLP to stabilize metric volatility. The following findings detail the operational trade-offs of these approaches, visualized in Figure 1. While lowering n-size thresholds significantly improves data reporting for small schools and historically underrepresented student groups, it also highlights an inherent tension between statistical reliability and sample size. Concurrently, while applying EBLP effectively reduces random noise and stabilizes test score metrics for smaller student populations, it does not translate into greater stability at the overall framework rating level.

Figure 1. Balancing transparency and fairness for small systems

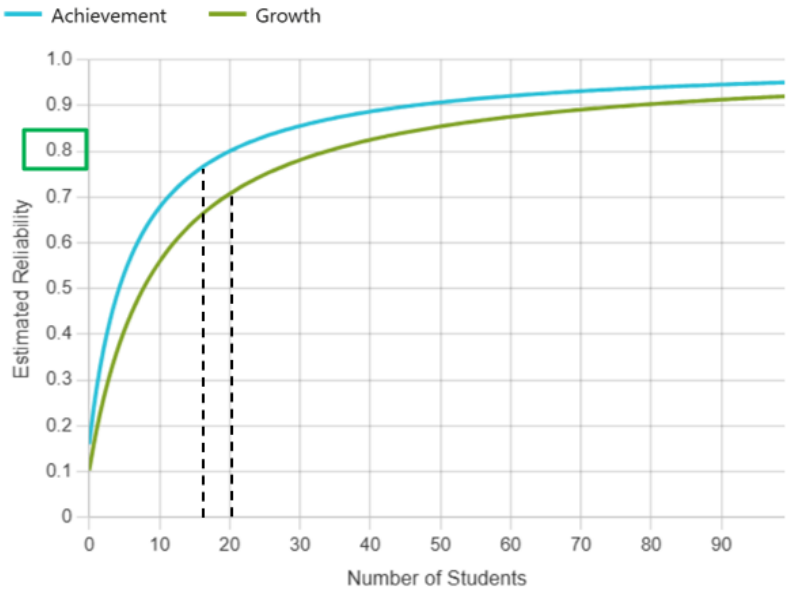


Lowering Student Count Thresholds Study Findings

CADRE's study found that lowering student count thresholds modestly increases overall reporting while significantly expanding data visibility for historically underreported student groups, though with statistical reliability trade-offs. Reducing the achievement N-size threshold from 16 to 10 expanded overall reportable data by 2–3% for schools and 4–5% for districts, with the largest proportional gains occurring at the school level for multilingual learners and students with IEPs. For growth data, lowering the N-size threshold from 20 to 10 yields even larger coverage gains, expanding overall reporting by 2–5% for schools and 6–7% for districts, while increasing disaggregated group reporting by 6% to 45% (most notably adding 498 elementary schools, or 45.4%, for students with IEPs).

However, these reporting gains introduce a separate statistical consideration: because reliability depends on both sample size and the intraclass correlation (ICC) — the share of variance in outcomes that is between-school rather than within-school — growth metrics and disaggregated subgroup data, which have consistently lower ICCs than achievement metrics or overall student populations, require larger sample sizes to reach comparable levels of reliability. This is distinct from the question of whether a given N-size threshold allows a unit to be reported at all. Figure 2 visualizes this relationship between reliability, ICC, and sample size. The dotted lines at 16 students and 20 students indicate the current minimum reporting thresholds and show how reliability estimates quickly drop off below these numbers for both achievement and growth. Reducing student count thresholds to 10 would produce substantially more volatile (i.e., less reliable) measures for small schools, districts, and student groups as compared to current thresholds.

Figure 2. Estimated Reliability by Sample Size and ICC Value



Note: Blue and green lines are based on plausible ICC values for Achievement and Growth. Reliability is measured on a scale of 0 to 1. A coefficient of 0.8 is generally considered adequate for reliable group-level reporting.

Measure Stabilization Study Findings

Analyzing three consecutive years of statewide test score data, CADRE's study found that EBLP successfully reduced metric volatility and produced significantly more accurate, stable results compared to traditional single-year measures. As seen in Table 1, which reports the stability of current mean scale score and stabilized EBLP estimates of achievement across schools from 2024 to 2025, these improvements were particularly pronounced for smaller schools, where small student sample sizes typically create high year-to-year score fluctuations.

Table 1. Stability of school mean versus EBLP scale scores across 2024-2025

				Mean Absolute Difference between 2024-2025 scores		Correlation between 2024-2025 scores	
Subj.	**EMH**	**School Size**	**Schools**	**Mean**	**EBLP**	**Mean**	**EBLP**
ELA	E	16-49	1423	6.34	4.17	0.93	0.97
		50-149	1582	4.30	3.26	0.94	0.96
		150+	673	3.23	2.71	0.98	0.98
	M	16-49	554	5.87	3.85	0.93	0.97
		50-149	671	4.42	3.13	0.95	0.98
		150+	688	3.28	2.73	0.97	0.98
	H	16-49	402	16.51	10.22	0.93	0.97
		50-149	372	11.31	7.67	0.97	0.98
		150+	434	8.04	6.14	0.98	0.99
Math	E	16-49	1424	6.16	4.10	0.91	0.96
		50-149	1560	4.07	3.21	0.94	0.96
		150+	702	3.02	2.68	0.98	0.98
	M	16-49	547	5.60	3.48	0.92	0.97
		50-149	664	3.88	2.70	0.96	0.98
		150+	704	2.81	2.32	0.98	0.99
	H	16-49	410	15.24	10.04	0.92	0.97
		50-149	370	10.40	7.10	0.97	0.98
		150+	442	7.49	5.70	0.98	0.99

Note: Sample includes all school by disaggregated group combinations with mean (current practice) and EBLP estimates in both 2024 and 2025. E=elementary (grades 3-5); M=middle school (grades 6-8); H=high school (grades 9-10).

However, the study also revealed a critical policy limitation: substituting single-year measures with EBLP-adjusted measures did not substantially improve the year-to-year stability of overall SPF ratings for small schools. As shown in Table 2, 72.6% of small schools had the same SPF Growth rating in 2024 as in 2025 using CDE's current methodology of calculating median SGPs, and 71.4% of small schools had the same SPF Growth rating between 2024 and 2025 using the EBLP stabilization methodology.

Table 2. Percent of Schools with the same SPF Growth rating in 2024 and 2025 using different measures

School Category	Median SGP	EBLP SGP
All schools	74.5%	74.4%
Large (n>=400)	80.5%	79.1%
Medium (150<=n<400)	70.3%	71.8%
Small (n<150)	72.6%	71.4%

Thus, while EBLP effectively filters out random noise at the individual metric level, its benefits do not necessarily carry over into stabilizing overall accountability classifications due to the structural cutoffs and boundary rules embedded within the broader SPF rating system.

Stakeholder feedback

As required by legislation, CDE shared findings from both studies over the course of 2025 and 2026 with the TAP, AWG, and Rural Education Council and solicited their recommendations for next steps. While the TAP, a group of education measurement experts, was consulted on the technical details of study methodologies and findings, CDE presented an abbreviated summary of key findings and facilitated discussions with the AWG and Rural Education Council.

TAP feedback

What is the TAP?

The Technical Advisory Panel for Longitudinal Growth (TAP) consists of district representatives, state, and national experts on longitudinal measurement of academic growth and framework measures for state accountability purposes, convened by the Commissioner of Education to provide recommendations to the State Board of Education. The TAP was created in accordance with the Education Accountability Act of 2009 (SB-09-163).

The TAP was consulted in multiple meetings over the course of 2025-2026 as the studies were under way. Recordings and summaries of conversations are publicly available and can be found on CDE's website. The TAP felt it best to maintain current thresholds (a minimum N of 16 for Achievement and 20 for Growth). A primary concern was that lowering these thresholds would increase volatility, reducing both the reliability and stability of results while creating a risk of less consistent year-to-year outcomes. Given the high-stakes implications of accountability systems, TAP members raised concerns regarding data reliability and student privacy, noting that the benefits of increased reporting did not clearly outweigh these risks. Ultimately, the TAP emphasized the value of the current thresholds, highlighting that existing levels are well-supported by data and prior rationale, providing stable and defensible results.

TAP members also expressed a preference to maintain the current calculation approach without statistical stabilization, viewing the perceived drawbacks as clearly outweighing the benefits. A primary and repeated concern was that stabilization offers limited added value relative to its methodological complexity, creating key risks such as reduced reproducibility and transparency in calculations, as well as operational disruption such as requiring renorming and training on new methodologies across the state. Conversely, feedback highlighted the strengths of the current approach, emphasizing that it preserves interpretability, consistency, and defensibility while maintaining a clear focus on actual scores and straightforward calculations.

AWG feedback

What is the AWG?

The Accountability Work Group (AWG) serves as an advisory group to CDE on accountability policy implementation and practice at the state, district and school levels, including representatives from small systems and rural sites across Colorado. While the AWG has been in operation for many years, H.B. 25-1278 formalized the group as a state advisory body.

CDE shared findings from the two studies and collected feedback from the AWG in May 2026. A recording of the presentation and summary can be found on CDE's website. To better serve small and rural districts, AWG members urged CDE to focus less on changes to framework calculations and shift towards meaningful, practice-based support. Some recommendations included:

- **Clear, Accessible Communication:** Provide high-level, clear summary guidance and maintain focus on secure data tools (such as the Secure Data Explorer) to help local leaders digest complex framework updates without feeling overwhelmed.
- **Differentiated Data Aggregation:** Explore creative ways to aggregate data for small sites—such as pooling data across combined grade bands (e.g., establishing a single K–8 performance rating for co-located elementary and middle schools with small N-counts) or lowering reporting thresholds while preserving accountability minimums.
- **Promoting Best Practices & Subgroup Supports:** Focus efforts on identifying and elevating successful practices from districts that excel in serving historically underserved student groups, ensuring that small systems receive targeted tools to monitor and support special education and disaggregated student populations.

Rural Education Council Feedback

What is the Rural Education Council?

The Rural Education Council, made up of rural superintendents and other organizations supporting rural areas, advises the Education Commissioner about the needs, challenges, and opportunities of rural districts.

Members of the Rural Education Council voiced similar concerns regarding the effects of introducing reliability issues into accountability frameworks with lower reporting thresholds. Council members felt that lowering the minimum N-size threshold for accountability could be overly problematic because unreliable data might unfairly trigger high-stakes consequences, such as placement on the five-year clock. While lower thresholds might be acceptable for non-punitive uses like grant funding access, the Rural Education Council pointed out that risks to schools outweigh the benefits of introducing these new calculation approaches.

Recommendations

Based on findings from CADRE's two studies as well as significant stakeholder input collected from the TAP, AWG, and Rural Education Council, CDE offers the following recommendations regarding changes to accountability in support of small systems.

- **Do not lower minimum reporting thresholds.** Setting a minimum N-size threshold for accountability reporting involves making tradeoffs between transparency and reliability, which ultimately stakeholders felt did not warrant a lower reporting threshold. Lower minimum N-size thresholds would enable reporting data for more units (particularly for disaggregated groups within schools and districts), but data reported for small groups or units are likely to be less reliable. In feedback conversations, stakeholders did not feel the tradeoff of improving reporting coverage for 2-3% of districts in achievement and 2-5% of districts in growth would be worth the lower reliability of data used for high-stakes decisions tied to performance frameworks. Similarly, a lowered reporting threshold raised concerns about student privacy for members of the TAP.
- **Do not adopt a stabilization methodology.** Adopting EBLP in state performance frameworks is not recommended due to significant operational complexity, administrative costs, and minimal practical benefit. Transitioning to EBLP would require CDE to rebuild its growth calculation infrastructure, re-norm growth measures, and retrain statewide stakeholders on complex modeling required for EBLP, a transition requiring substantial resource updates and an additional fiscal note. Beyond these fiscal and operational hurdles, EBLP could compromise framework transparency and public trust by holding schools accountable for imputed statistical estimations rather than observed student performance.

- **Support small systems in other facets of the accountability system.** To effectively support small school systems, CDE should look to other components of the state accountability system outside of performance frameworks (see Figure 3). Based on stakeholder feedback, the department recommends that the state support public reporting for small systems and promote internal data tools (such as the Secure Data Explorer) where rural leaders can track student performance privately. CDE should also highlight small schools and districts that successfully support historically underserved student groups, sharing their strategies as practical peer-to-peer support, planning, and intervention playbooks for other rural systems. Similarly, CDE should design more customized guidance and technical assistance specifically to small-school operating realities rather than retrofitting large-district practices.

Figure 3. State Accountability Elements



Conclusion

Based on findings from two studies exploring lowering minimum N-size for reporting and stabilization techniques for small systems, along with feedback from the TAP, AWG, and Rural Education Council, CDE recommends keeping the current framework calculation methodology and instead supporting small systems through other aspects of the accountability system.

Overall, reducing the minimum N-size to 10 students for achievement and growth yields modest increases for overall school populations, but a substantial increase in reported data for disaggregated student subgroups (e.g., race/ethnicity, English language learners, students with disabilities). Statistical reliability directly scales with sample size. Smaller student groups produce less reliable aggregate scores, though lowering N-size increases visibility for historically underrepresented student populations. Stakeholders felt the public benefit of reporting subgroup data did not outweigh the risk of making high-stakes decisions based on lower reliability metrics.

EBLP produces more accurate and stable test score metrics compared to traditional, unadjusted single-year averages and the stabilization effect is strongest in small schools, where small student samples make traditional single-year scores most vulnerable to annual fluctuation. However, substituting EBLP-adjusted measures for single-year measures did not increase the overall stability of SPF Growth Ratings. Additionally, EBLP would add a layer of opacity to CDE's accountability calculations, and stakeholders felt the tradeoff in complexity did not outweigh the benefits this methodology would bring to SPFs.

References

Accountability, A. S. (2024). *1241: Accountability, Accreditation, Student Performance and Resource Inequity Task Force Report.* Denver: Colorado Department of Education. Retrieved from https://resources.finalsite.net/images/v1780335732/cdestatecous/xu2urfr9lp13udzrau78/FullReport_1241TaskForce_Final.pdf

Shear, B. R., Whitfield, E., & Mangan, D. (2026). *Investigation of EBLP to Stabilize School Test Score Metrics in Colorado.* Boulder: Center for Assessment, Design, Research and Evaluation (CADRE) University of Colorado Boulder. Retrieved from https://www.colorado.edu/cadre/media/385

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: Center for Assessment, Design, Research and Evaluation (CADRE) University of Colorado Boulder. Retrieved from https://www.colorado.edu/cadre/media/378

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....4

Purpose....5

Background....6

Current Colorado Minimum N-Size Thresholds....6
Aggregate Achievement and Growth Measures....6
Reliability of School Aggregate Test Score Metrics....8

Data....9

Methods....10

Results....12

Sample Size Impacts on Available Data....12
ICC Estimates....14
Reliability Estimates....16

Discussion....18

Appendix....20

References....24

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.

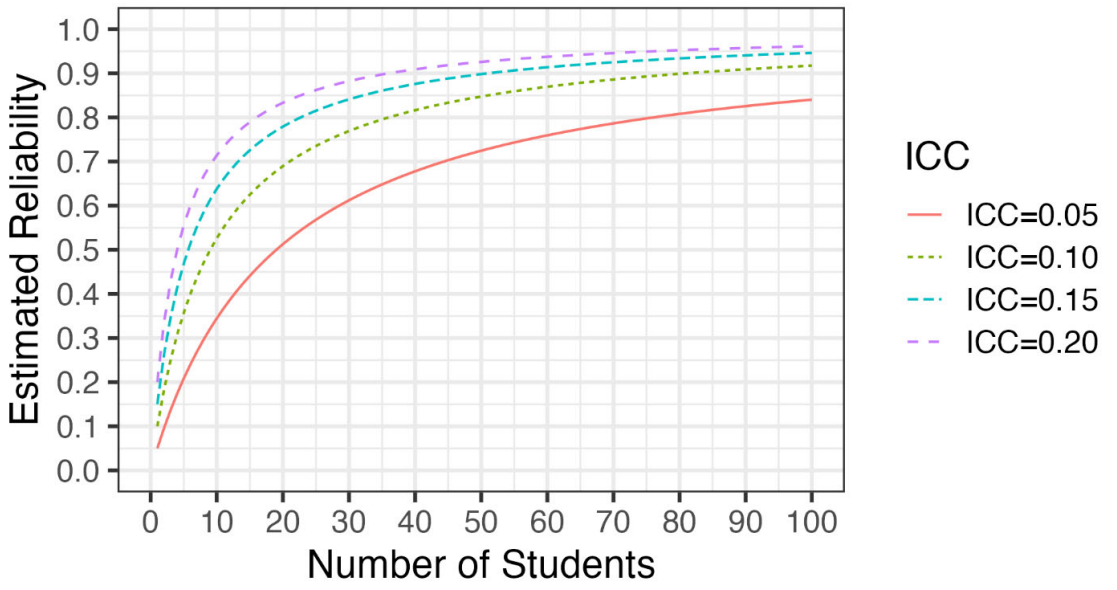
Table 5: Implied Reliability by Sample Size and ICC Value

	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

CADRE

MAY 2026

Investigation of EBLP to Stabilize School Test Score Metrics in Colorado

Benjamin R. Shear[1], Erik Whitfield[1], Daniel Mangan[2]

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

[1] Center for Assessment, Design, Research & Evaluation (CADRE) School of Education
University of Colorado Boulder

[2] Colorado Department 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 Marie Huchton and Anne Laesecke at the Colorado Department of Education.

Suggested Citation

Shear, B. R., Whitfield, E. W., & Mangan, D. (2026). *Investigation of EBLP to stabilize school test score metrics in Colorado*. Boulder, CO: The Center for Assessment, Design, Research and Evaluation (CADRE), University of Colorado Boulder.

Please send correspondence concerning this report to Benjamin Shear at benjamin.shear@colorado.edu.

Table of Contents

Introduction 4

Background 5

Empirical Best Linear Prediction (EBLP) 6

Purpose 7

Data 7

Methods 8

Results 10

Distribution of Achievement and Growth Sample Sizes 10

Distribution of Mean Scale Scores, Mean SGP, and Median SGP Across Schools 11

Stability of Direct versus EBLP Stabilized Estimates 13

Accuracy of Direct versus EBLP Stabilized Estimates 16

Comparison of SPF Ratings 17

Discussion 20

References 23

Introduction

Standardized test score metrics are a core component of Colorado's school accountability ratings. The Colorado 1241 Task Force[1] report notes a perceived concern that test scores and other metrics, including the School Performance Framework (SPF) ratings derived from them, can experience high volatility, particularly for schools and districts serving small student populations. In response, the Task Force made the following recommendation (#4): "Explore best practices and monitor the accountability system to identify and reduce issues of volatility that impact schools and districts with small student populations" (p. 27). The current report presents exploratory analyses intended to support this recommendation. First, we provide a more formal framework for considering volatility of test score metrics, which is not precisely defined in the Task Force Report. Second, we review a recently developed statistical approach to reduce annual volatility of aggregate test score metrics. Third, we apply this method to Colorado test score data to demonstrate characteristics of the method and provide empirical evidence that the Colorado Department of Education (CDE) can use to inform how these methods could impact Colorado SPF ratings.[2] Finally, we summarize the results and discuss important considerations related to applying statistical stabilization methods within test score-based school accountability systems.

[1] https://ed.cde.state.co.us/accountability/stakeholders/accountability-task-force

[2] The analyses in this report focus on school-level test score metrics and SPF ratings. However, similar concerns about volatility could apply to district-level test score metrics and District Performance Framework (DPF) ratings and the same methods could be applied to district-level data. Consideration of empirical results for districts is beyond the scope of the present analyses.

Background

Standardized test results are a core component of Colorado's School Performance Framework (SPF). At the school level, aggregate test score metrics (mean scale scores and median student growth percentiles [SGP] for all students and for specific disaggregated student groups) summarizing achievement and growth are incorporated into accountability determinations. The Colorado 1241 Task Force highlighted a concern that these metrics—and SPF ratings derived from them—can be highly volatile, especially for schools and districts serving small numbers of students. Although "volatility" is not always defined formally in policy discussions, it generally refers to year-to-year changes in reported school or district test score metrics that can occur even when underlying performance is relatively stable. The 1241 Task Force report references three mechanisms that may contribute to this volatility (p. 23), which are summarized below. We note that while the terminology below is not used in the Task Force report, these mechanisms are described.

- **Insufficient sample sizes for reporting.** When the number of tested students falls below reporting rules (N=16 for achievement and N=20 for growth), results may be suppressed or incomplete, limiting the indicators and sub-indicators included in reporting and ratings.
- **Threshold effects.** When tested student counts are near reporting thresholds, small enrollment changes year over year can determine whether results are reported at all, producing differences in what, if any, information is available across years.
- **Sampling variation in small groups.** Even when reporting thresholds are met, aggregate results based on small numbers of students are less stable than results based on larger numbers of students. In small schools (or small student groups within larger schools) the performance of a few students can more substantially affect a mean scale score or median SGP. As a result, some portion of year-to-year change reflects idiosyncratic fluctuation of test score performance for small numbers of students rather than meaningful changes in student learning.

This report focuses on the third issue: settings where results are reportable, but small tested student counts still yield comparatively large year-to-year variation. This is a widely recognized challenge in test-based accountability systems. Because annual aggregate test score metrics tend to be more volatile when group sizes are small, the impacts of this volatility are expected to be unevenly distributed across schools due to differences in enrollment size.

Empirical Best Linear Prediction (EBLP)

One recently developed approach to addressing volatility in aggregate test scores or growth scores involves *stabilizing* averages using a statistical framework known as Empirical Best Linear Prediction (EBLP; Castellano et al., 2023; Lockwood et al., 2022).[3] The EBLP approach produces a stabilized test score metric by combining a school's current-year average with additional information (primarily the school's own results from prior years) in a way that depends on how precisely the current-year value is estimated. The EBLP approach is not intended to replace the underlying student test scores or growth scores; rather, it is a way to report more stable year-to-year results when aggregating student-level test scores or growth scores at the school level.

Conceptually, stabilized metrics produced by the EBLP approach can be thought of as weighted averages across years (Castellano et al., 2023; McCaffrey & Castellano, 2021). When a school's current-year mean is based on many tested students, it is relatively precise and receives most of the weight; in this case, the stabilized estimate will be close to the current-year simple average. With smaller sample sizes the simple average will be less precise, and prior-year information will receive more weight. The Colorado multi-year framework calculations, which compute averages or medians across three years of data for small schools, are equivalent to calculating a weighted average across three years with weights proportional to the number of tested students within each year. In contrast, the EBLP stabilized estimates apply weights to each year that are designed to maximize the accuracy of the weighted average as a current year estimate for each school. The method thus tends to produce the largest stabilization effects in the smallest schools and student groups while leaving estimates for large schools largely unchanged.

The peer-reviewed literature motivating this approach emphasizes two linked goals: improving accuracy of aggregate growth estimates (in the sense of being closer to the underlying signal of interest) and improving stability (reducing year-to-year fluctuation), with the largest gains expected when the reporting group is small. Throughout the report we refer to aggregate test score metrics produced using the EBLP approach as "stabilized estimates" or "EBLP estimates." In contrast, we refer to averages or medians computed using only the current year data as "simple averages," "simple medians," or "direct estimates" to contrast them with the stabilized estimates. We note that at the current time, the EBLP approach can only be used to produce stabilized averages or means, but not stabilized medians. Colorado uses the median rather than the mean to summarize SGP at the school and district level, and we discuss the implications of this in our methods section below.

The EBLP framework is currently used to produce school growth scores for California schools, local education agencies (LEA), and student groups in grades 4-8. The EBLP approach was developed specifically to address concerns about volatility of average school and LEA growth scores across years. School and LEA growth scores in California are reported for informational purposes but are not currently used as part of "Accountability support determinations."[4] The following specifications are used to produce growth scores for each school, LEA, or student group. [5] First, a growth score for each enrolled student is computed using each student's current test score and prior year test scores. Second, for units enrolling 500 or more

[3] While some sources use the term "average" to refer either to the arithmetic mean or to the median of a dataset, we use the term "average" to refer only to the arithmetic mean. We thus use the terms "average" and "mean" interchangeably.

[4] https://www.cde.ca.gov/ta/ac/growthmodelfaq.asp

[5] For additional details see: https://www.cde.ca.gov/ta/ac/acctgrowthmod.asp

students with growth scores, a simple average based on current year student growth scores is reported. For units enrolling 500 or fewer students with growth scores, the EBLP approach is used to calculate a stabilized estimate of average growth scores, using the two most recent years of growth scores for the school. This hybrid rule is intended to balance "the need for score stability from year to year with the need to show changes in true growth as soon as it can be done with reasonable confidence."[6]

Purpose

This report is intended to evaluate how using the EBLP stabilization approach would impact school-level achievement and growth metrics in Colorado schools. Prior research (Castellano et al., 2023; Lockwood et al., 2022) and technical reports prepared for the California Department of Education[7] document that use of the EBLP approach can produce more accurate and more stable year-to-year school growth measures in other states and contexts, particularly for smaller schools or student groups. In this report we address the following two research questions:

1. How much does the EBLP method increase accuracy and reduce volatility of mean scale scores and mean SGP among Colorado schools?
2. Does use of EBLP stabilized mean scale scores and mean SGP reduce the volatility of SPF ratings?

Data

We use three years of statewide assessment results spanning 2022-23 through 2024-25 provided by CDE. The data consist of student-level longitudinal records. Specifically, we include student test scores in grades 3-11 and student SGP in grades 4-11 across three years (2022-23, 2023-24, and 2024-25) in two subjects: English Language Arts (ELA) and Mathematics.[8] The test scores and SGP are based on the Colorado Measures of Academic Success (CMAS) for grades 3-8 in both subjects; PSAT 9, PSAT 10, and SAT for grades 9, 10, and 11 respectively. We include Colorado Spanish Language Arts (CSLA) assessment results as part of the ELA data for students tested in Spanish in grades 3-5, consistent with CDE accountability business rules. We include Dynamic Learning Maps (DLM) alternate assessment results for students with significant cognitive disabilities, also consistent with CDE accountability practices.

While our analytical sample is designed to replicate CDE's official accountability calculations as closely as possible, some results may differ slightly from officially reported values due to differences in data processing, application of business rules, or other technical specifications.

[6] See FAQ "Why do you weight the scores for fewer than 500 students at a school, LEA or within a student group?" at: https://www.cde.ca.gov/ta/ac/growthmodelfaq.asp

[7] See state board memorandum of February 2021 at https://www.cde.ca.gov/be/pn/im/documents/feb21memoamard02.docx.

[8] Student SGP for Grade 9 ELA are not included, as Colorado does not calculate or report these operationally due to differences between the 8th grade CMAS ELA tests and the PSAT 9 Reading & Writing tests.

Methods

This analysis employs the EBLP methodology described above to stabilize school performance measures by pooling information across multiple years of data. As described above, the EBLP method calculates a weighted average of performance data from multiple years (the current focal year and at least one prior year) such that the mean squared error is minimized in estimating school-level performance for the current focal reporting year. The EBLP procedure thus requires at least two consecutive years of student-level data to produce stabilized estimates. For the 2024-25 school year, we calculate EBLP estimates by combining student performance data from 2023-24 and 2024-25. For the 2023-24 school year, we calculate EBLP estimates by combining student performance data from 2022-23 and 2023-24. The direct estimates are calculated using only data from the current focal year.[9]

We compute direct and stabilized estimates for each school level (elementary, middle, or high school, denoted by "EMH") by subject by disaggregated group by year combination. The school EMH levels are the same CDE designations used for SPF calculations; in general, this combines results for grades 3-5 for elementary, grades 6-8 for middle, and grades 9-11 for high school. Consistent with CDE accountability business rules, we include all available scale scores or SGP associated with each school, with the following exceptions. When computing mean scale scores at the high school level, we exclude grade 11 SAT scores, because these are used in the Postsecondary and Workforce Readiness (PWR) indicator. When computing mean or median SGP at the high school level, we exclude grade 9 SGP for ELA (which are based on the PSAT 9 Reading & Writing tests) because CDE does not compute or report these operationally. We compute results for each of the following disaggregated student groups for each school and subject: all students, minority students[10], students with an IEP, students eligible for free or reduced-price lunch (FRL), and multilingual learners (ML).

We follow CDE business rules for accountability calculations to produce school-level indicators (mean test scores or mean and median SGP). These rules establish minimum thresholds of 16 students for achievement indicators and 20 students for growth indicators. For achievement indicators measured by scale scores, we calculate a direct school mean when the school has at least 16 valid scale scores in the current year and we produce an EBLP stabilized estimate for any school meeting two criteria: (1) at least 16 scale scores in the current reporting year, and (2) at least 1 scale score in the prior year. For growth indicators measured by SGP, we calculate a direct school mean when the school has at least 20 valid SGP values in the current year and we produce an EBLP stabilized estimate for any school with at least 20 SGP values in the current year and at least 1 SGP value in the prior year. We apply the same rules to calculate direct and stabilized estimates for disaggregated student groups within each school. When computing the stabilized estimates for a disaggregated student group within a school, the EBLP estimate is a weighted average of results for only that disaggregated group within that school.

[9] Note that while the analyses reported here rely only on data reported for three years (2022-23, 2023-24, and 2024-25), the student-level SGP in these data were computed using additional years of data before being provided by CDE. For example, the 2022-23 SGP were computed based on student prior scores from 2021-22, even though 2021-22 data are not directly included in our analyses.

[10] Colorado includes all students with a reported race/ethnicity designations other than White as "minority" students for accountability reporting purposes. These race/ethnicity groups are: American Indian or Alaska Native, Asian, Black or African American, Hispanic or Latino, Native Hawaiian or Other Pacific Islander, or Two or More Races.

All comparisons between direct and EBLP estimates are restricted to the sample of schools that have EBLP estimates (and therefore also direct estimates) available in both 2023-24 and 2024-25. This ensures that differences in year-to-year stability reflect changes in estimation approach rather than changes in the composition of schools with reportable data. As noted above, EBLP estimates can only be constructed for means, not medians. We produce both direct mean and direct median SGP and compare these to stabilized mean SGP to enable comparisons.

To evaluate the stability of direct and EBLP estimates, we use two different metrics. First, we compute the mean absolute difference (MAD) of each indicator across the two years of data. This statistic quantifies how much each school's direct or EBLP achievement or growth indicator changes from one year to the next, on average. Lower values indicate less change (and thus greater stability) across years. Second, we compute the Pearson correlation coefficient of school indicators across the two years of data. The correlation is regularly used to assess the stability of aggregate school indicators from year to year and quantifies how stable the relative value of each school's indicator remains (relative to other schools) from one year to the next. Higher values of the correlation indicate greater stability across years.

To evaluate the accuracy of direct and EBLP estimates, we compute the mean accuracy ratio of the direct versus EBLP estimates using Equation 6 from Castellano et al. (2023). This statistic compares the ratio of mean squared error for the direct versus EBLP estimates across schools. An accuracy ratio of 1.0 for a school indicates the direct and EBLP estimates are equally accurate, whereas accuracy ratios greater than 1.0 indicate the EBLP estimate is more accurate.

All analyses were carried out using the R statistical software (R Core Team, 2025). We produce EBLP estimates using the *schoolgrowth* package in the R programming language (Lockwood & Castellano, 2022). To implement this, we run the schoolgrowth() function separately for each school EMH by subject by disaggregated group by year combination.

Results

Distribution of Achievement and Growth Sample Sizes

Table 1 reports the distribution of student sample sizes across schools for achievement results. We present two views of this information. The first summarizes the number of all tested students across schools. The second summarizes the number of students in each disaggregated student group across schools; thus each school can contribute more than one observation to this section if they have students in more than one disaggregated group. Elementary, middle, and high schools have on average 164, 266, and 285 total tested students in ELA and 166, 270, and 287 in math, based on just under 970, 530, and 370 schools, respectively. There are fewer high schools than middle and elementary schools, but high schools span the widest range of enrollment sizes. Breaking schools into disaggregated student groups increases the number of reported units and decreases the typical number of tested students.

Table 1: Descriptive statistics of school sample sizes for achievement metrics in 2025

Sample	EMH	Subject	Schools	Min	Mean	Median	Max
All Students	*E*	*ELA*	*967*	*18*	*164*	*159*	*911*
	E	*MATH*	*965*	*18*	*166*	*162*	*918*
	M	*ELA*	*529*	*16*	*266*	*210*	*1231*
	M	*MATH*	*527*	*17*	*270*	*212*	*1232*
	H	*ELA*	*369*	*16*	*285*	*178*	*1482*
	H	*MATH*	*370*	*16*	*287*	*179*	*1491*
All Disaggregated Groups	*E*	*ELA*	*3678*	*16*	*91*	*71*	*911*
	E	*MATH*	*3686*	*16*	*92*	*72*	*918*
	M	*ELA*	*1913*	*16*	*154*	*96*	*1231*
	M	*MATH*	*1915*	*16*	*157*	*98*	*1232*
	H	*ELA*	*1208*	*16*	*170*	*87*	*1482*
	H	*MATH*	*1222*	*16*	*171*	*86*	*1491*

Note. E=elementary (grades 3-5); M=middle school (grades 6-8); H=high school (grades 9-10). The “All Students” sample reports results based on all tested students in each school. The “All Disaggregated Groups” sample reports results across school by disaggregated group combination, including the results for all students in a school. This table summarizes 2025 data for schools in the analytic sample, which includes schools with relevant data in 2024 and 2025.

Table 2 reports analogous information to Table 1 but summarizing the number of students with growth (SGP) data. Growth calculations require test scores from consecutive years, which reduces the number of students and thus reduces the number of schools with reportable data. Elementary, middle, and high schools, have growth scores for 102, 251, and 272 students in ELA, and 105, 253, and 373 in math, on average. There is a notable difference in sample sizes (number of schools with data) between subjects at the high school level because growth scores do not include SGP values for 9th grade ELA, thus reducing the number of high schools with sufficient data for reporting. Breaking schools into disaggregated student groups increases the number of reported observations and decreases the typical number of student observations.

Table 2: Descriptive statistics of school sample sizes for growth metrics in 2025

Sample	EMH	Subject	Schools	Min	Mean	Median	Max
All Students	*E*	*ELA*	*922*	*20*	*102*	*96*	*644*
	E	*MATH*	*917*	*20*	*105*	*98*	*643*
	M	*ELA*	*514*	*20*	*251*	*200*	*1169*
	M	*MATH*	*513*	*20*	*253*	*202*	*1166*
	H	*ELA*	*345*	*20*	*272*	*187*	*1390*
	H	*MATH*	*371*	*21*	*373*	*239*	*1833*
All Disaggregated Groups	*E*	*ELA*	*2569*	*20*	*70*	*61*	*644*
	E	*MATH*	*2629*	*20*	*73*	*64*	*643*
	M	*ELA*	*1733*	*20*	*151*	*100*	*1169*
	M	*MATH*	*1736*	*20*	*153*	*100*	*1166*
	H	*ELA*	*1012*	*20*	*172*	*98*	*1390*
	H	*MATH*	*1207*	*20*	*223*	*111*	*1833*

Note. E=elementary (grades 4-5); M=middle school (grades 6-8); H=high school (grades 9-11, excluding grade 9 ELA). The "All Students" sample reports results based on all students with SGP in each school. The "All Disaggregated Groups" sample reports results across school by disaggregated group combination, including the results for all students in a school. This table summarizes 2025 data for schools in the analytic sample, which includes schools with relevant data in 2024 and 2025.

Distribution of Mean Scale Scores, Mean SGP, and Median SGP Across Schools

Table 3 summarizes the distribution of simple (direct) mean scale scores across schools while Table 4 summarizes the distribution of simple mean and median SGP across schools. These tables again summarize 2025 data for schools in the analytic sample (schools with data in both 2024 and 2025). Results are similar when summarizing 2024 data. Each table again reports results when aggregating across all students in each school ("All Students") or across separate disaggregated student groups in each school ("All Disaggregated Groups).

The mean scale scores summarized in Table 3 vary considerably across schools. The distribution of mean scale scores is similar in math and ELA, with slightly lower means and larger standard deviations (SD) overall for the Disaggregated Groups results. The lower mean scale scores and larger SD for high schools relative to elementary and middle schools results from differences between the PSAT 9/10 and CMAS reporting scales.

Table 3: Distribution of school mean scale scores in 2025

Sample	EMH	Subject	Schools	Mean	SD	Min	Max
All Students	*E*	*ELA*	*967*	*740.5*	*17.5*	*693.5*	*791.6*
	E	*MATH*	*965*	*737.1*	*17.3*	*696.6*	*788.1*
	M	*ELA*	*529*	*742.2*	*17.0*	*684.7*	*788.4*
	M	*MATH*	*527*	*731.4*	*17.3*	*688.8*	*785.0*
	H	*ELA*	*369*	*455.6*	*46.6*	*313.0*	*577.9*
	H	*MATH*	*370*	*429.9*	*46.6*	*290.1*	*561.4*
All Disaggregated Groups	*E*	*ELA*	*3678*	*728.6*	*19.8*	*677.9*	*792.9*
	E	*MATH*	*3686*	*727.2*	*18.1*	*683.1*	*793.3*
	M	*ELA*	*1913*	*730.8*	*18.9*	*681.9*	*788.4*
	M	*MATH*	*1915*	*721.5*	*17.8*	*682.3*	*785.8*
	H	*ELA*	*1208*	*423.2*	*52.7*	*295.0*	*584.6*
	H	*MATH*	*1222*	*401.4*	*51.1*	*275.9*	*569.8*

Note. E=elementary (grades 3-5); M=middle school (grades 6-8); H=high school (grades 9-10); SD=standard deviation. The "All Students" sample reports results based on all students with SGP in each school. The "All Disaggregated Groups" sample reports results across school by disaggregated group combination, including the results for all students in a school. This table summarizes 2025 data for schools in the analytic sample, which includes schools with relevant data in 2024 and 2025.

Comparison of the median and mean SGP in Table 4 reveals some important characteristics of these two different metrics. While the means of the school median and school mean SGP are similar, the SD and range of the school median SGP are consistently greater than the SD and range of the school mean SGP. For example, the SD of median SGP across schools is about 40% larger than the SD of the mean SGP in each row of Table 4, while the range of the median SGP is about 30% larger. Thus, school median SGP are more variable than school mean SGP within years. This becomes relevant when we consider implications for constructing SPF ratings from these two different metrics using normative thresholds, as described in more detail below.

Table 4: Distribution of school median and mean SGP in 2025

				Median SGP				Mean SGP			
Sample	**EMH**	**Subject**	**Schools**	**Mn**	**SD**	**Min**	**Max**	**Mn**	**SD**	**Min**	**Max**
All Students	*E*	*ELA*	*922*	*50.6*	*10.9*	*17.5*	*81.0*	*50.2*	*7.6*	*22.2*	*71.1*
	E	*MATH*	*917*	*50.7*	*13.1*	*10.0*	*93.0*	*50.2*	*9.4*	*16.7*	*85.9*
	M	*ELA*	*514*	*49.9*	*10.2*	*19.0*	*83.5*	*49.9*	*7.2*	*26.4*	*74.9*
	M	*MATH*	*513*	*49.3*	*10.9*	*15.0*	*82.5*	*49.4*	*7.7*	*24.9*	*74.7*
	H	*ELA*	*345*	*47.8*	*8.4*	*23.0*	*74.0*	*48.3*	*5.6*	*32.5*	*64.0*
	H	*MATH*	*371*	*48.1*	*9.7*	*17.0*	*76.5*	*48.5*	*6.8*	*23.6*	*69.6*
All Disaggregated Groups	*E*	*ELA*	*2569*	*48.5*	*10.9*	*13.0*	*89.0*	*48.7*	*7.6*	*21.0*	*74.0*
	E	*MATH*	*2629*	*49.2*	*13.2*	*10.0*	*93.0*	*49.1*	*9.3*	*16.7*	*85.9*
	M	*ELA*	*1733*	*48.4*	*10.4*	*19.0*	*87.5*	*48.8*	*7.2*	*23.3*	*74.9*
	M	*MATH*	*1736*	*47.5*	*10.8*	*15.0*	*88.0*	*48.0*	*7.7*	*23.1*	*75.5*
	H	*ELA*	*1012*	*45.5*	*8.7*	*16.0*	*74.0*	*46.6*	*6.0*	*21.1*	*68.0*
	H	*MATH*	*1207*	*45.5*	*10.3*	*14.5*	*77.0*	*46.5*	*7.2*	*23.3*	*69.6*

Note. E=elementary (grades 3-5); M=middle school (grades 6-8); H=high school (grades 9-10); SD=standard deviation. The "All Students" sample reports results based on all students with SGP in each school. The "All Disaggregated Groups" sample reports results across school by disaggregated group combination, including the results for all students in a school. This table summarizes 2025 data for schools in the analytic sample, which includes schools with relevant data in 2024 and 2025.

Stability of Direct versus EBLP Stabilized Estimates

Table 5 reports the stability of direct and stabilized EBLP estimates of achievement across schools from 2024 to 2025, separately by subject and sample size. Results are summarized separately for estimates based on 16-49 students, 50-149 students, or 150 or more students. This table summarizes the All Disaggregated Groups samples that include results for all disaggregated groups at each school. As anticipated, the EBLP estimates are more stable across years from 2024 to 2025, with larger differences between the direct and EBLP values estimated from smaller sample sizes. For example, the MAD of direct mean estimates for elementary school ELA scores is 6.34 for schools with 16-49 students, 4.30 for schools with 50-149 students, and 3.23 for schools with 150+ students. These values are 52%, 32%, and 19% larger than the corresponding values for the EBLP estimates. This indicates that, on average, the direct school mean scale scores change more from year to year on average than do the EBLP estimates. The correlation between EBLP estimates is also greater than the correlation of the direct estimates.

Table 5: Stability of school mean (direct) versus EBLP scale scores

				MAD		Correlation	
Subject	**EMH**	**Size**	**Schools**	**Mean**	**EBLP**	**Mean**	**EBLP**
ELA	*E*	*16-49*	*1423*	*6.34*	*4.17*	*0.93*	*0.97*
ELA	*E*	*50-149*	*1582*	*4.30*	*3.26*	*0.94*	*0.96*
ELA	*E*	*150+*	*673*	*3.23*	*2.71*	*0.98*	*0.98*
ELA	*M*	*16-49*	*554*	*5.87*	*3.85*	*0.93*	*0.97*
ELA	*M*	*50-149*	*671*	*4.42*	*3.13*	*0.95*	*0.98*
ELA	*M*	*150+*	*688*	*3.28*	*2.73*	*0.97*	*0.98*
ELA	*H*	*16-49*	*402*	*16.51*	*10.22*	*0.93*	*0.97*
ELA	*H*	*50-149*	*372*	*11.31*	*7.67*	*0.97*	*0.98*
ELA	*H*	*150+*	*434*	*8.04*	*6.14*	*0.98*	*0.99*
MATH	*E*	*16-49*	*1424*	*6.16*	*4.10*	*0.91*	*0.96*
MATH	*E*	*50-149*	*1560*	*4.07*	*3.21*	*0.94*	*0.96*
MATH	*E*	*150+*	*702*	*3.02*	*2.68*	*0.98*	*0.98*
MATH	*M*	*16-49*	*547*	*5.60*	*3.48*	*0.92*	*0.97*
MATH	*M*	*50-149*	*664*	*3.88*	*2.70*	*0.96*	*0.98*
MATH	*M*	*150+*	*704*	*2.81*	*2.32*	*0.98*	*0.99*
MATH	*H*	*16-49*	*410*	*15.24*	*10.04*	*0.92*	*0.97*
MATH	*H*	*50-149*	*370*	*10.40*	*7.10*	*0.97*	*0.98*
MATH	*H*	*150+*	*442*	*7.49*	*5.70*	*0.98*	*0.99*

Note. MAD=mean absolute difference. EBLP=empirical best linear predictor. Sample includes all school by disaggregated group combinations with direct and EBLP estimates in both 2024 and 2025. E=elementary (grades 3-5); M=middle school (grades 6-8); H=high school (grades 9-10).

Table 6 reports the stability of direct mean and median SGP relative to stabilized EBLP mean SGP estimates from 2024 to 2025, also by subject and sample size for the All Student Group samples that include results for all disaggregated groups at each school. The MAD of direct median SGP are between 36% and 55% larger than the MAD of direct mean SGP, while the MAD of direct mean SGP are between 9% and 84% larger than the MAD of EBLP mean SGP, depending on sample size, EMH level, and subject. The cross-year correlation of direct mean SGP ranges from 2% smaller to 28% larger than the correlation of direct median SGP. The cross-year correlation of EBLP mean estimates ranges from 8% larger to 78% larger relative to the cross-year correlation of direct mean SGP estimates. In sum, direct mean SGP are more stable across years than direct median SGP, while EBLP mean SGP estimates are more stable across years than direct mean SGP. However, the differences between direct mean and median SGP cross-year correlations are relatively constant across sample sizes, while the differences between the direct mean and EBLP cross-year correlations are most pronounced for schools with smaller sample sizes. For example, while the correlation of direct mean SGP across years is higher than the correlation of direct median SGP, the correlation of direct mean SGP is also higher for schools with larger samples than for schools with smaller samples. In contrast, the correlation of EBLP mean SGP across years are similar across schools of different sample sizes (within EMH and subject combinations). The results are consistent with two theoretical expectations. First, that mean SGP tend to be more stable across years than median SGP, and

this difference is relatively consistent across school sample sizes. Second, the EBLP method increases stability of school-level growth estimates and the increases in stability are greatest for smaller schools. Although this second phenomenon is also observed with the school mean scale score data in Table 5, the result is less pronounced in Table 5 because direct estimates of mean scale scores have much higher stability to begin with relative to direct estimates of school mean SGP.

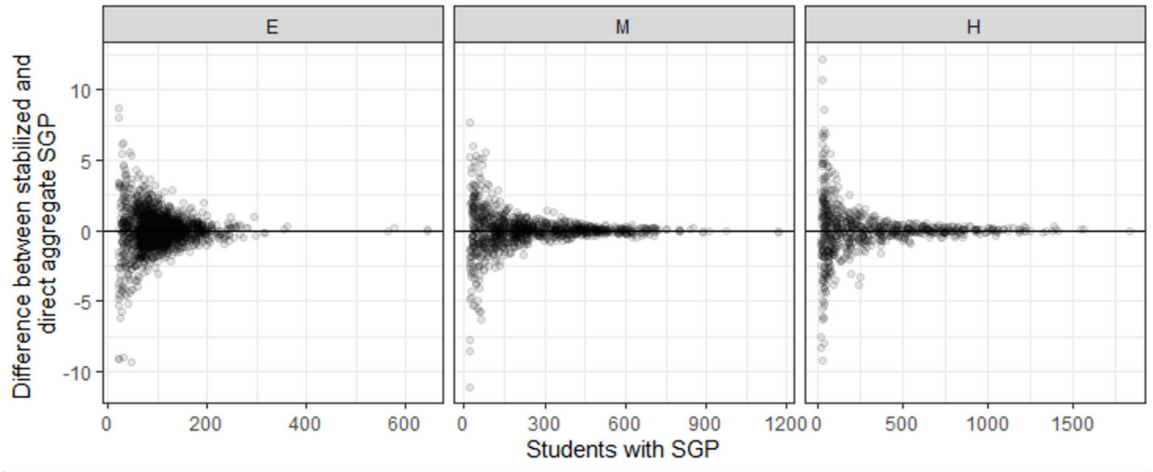
Table 6: Stability of school median, mean, and EBLP SGP

				MAD			Correlation		
Subject	**EMH**	**Size**	**N**	**Median**	**Mean**	**EBLP**	**Median**	**Mean**	**EBLP**
ELA	*E*	*20-49*	*1023*	*10.83*	*7.11*	*4.41*	*0.35*	*0.41*	*0.71*
ELA	*E*	*50-149*	*1411*	*8.64*	*5.70*	*4.28*	*0.48*	*0.52*	*0.69*
ELA	*E*	*150+*	*135*	*7.91*	*5.19*	*4.67*	*0.45*	*0.49*	*0.57*
ELA	*M*	*20-49*	*465*	*10.75*	*7.22*	*4.34*	*0.34*	*0.41*	*0.71*
ELA	*M*	*50-149*	*646*	*8.28*	*5.64*	*4.21*	*0.48*	*0.52*	*0.70*
ELA	*M*	*150+*	*622*	*6.23*	*4.22*	*3.75*	*0.62*	*0.64*	*0.70*
ELA	*H*	*20-49*	*297*	*9.31*	*6.59*	*3.57*	*0.46*	*0.46*	*0.81*
ELA	*H*	*50-149*	*338*	*6.16*	*4.10*	*2.40*	*0.59*	*0.65*	*0.86*
ELA	*H*	*150+*	*377*	*4.49*	*3.00*	*1.95*	*0.68*	*0.71*	*0.88*
MATH	*E*	*20-49*	*979*	*12.20*	*8.20*	*5.79*	*0.38*	*0.43*	*0.63*
MATH	*E*	*50-149*	*1503*	*9.53*	*6.49*	*5.52*	*0.53*	*0.56*	*0.65*
MATH	*E*	*150+*	*147*	*7.70*	*5.65*	*5.20*	*0.61*	*0.59*	*0.64*
MATH	*M*	*20-49*	*465*	*10.31*	*6.90*	*4.09*	*0.44*	*0.50*	*0.77*
MATH	*M*	*50-149*	*644*	*7.91*	*5.35*	*4.02*	*0.58*	*0.62*	*0.76*
MATH	*M*	*150+*	*627*	*5.96*	*4.08*	*3.53*	*0.68*	*0.70*	*0.77*
MATH	*H*	*20-49*	*309*	*11.53*	*7.42*	*4.08*	*0.39*	*0.50*	*0.81*
MATH	*H*	*50-149*	*368*	*7.19*	*5.02*	*3.13*	*0.69*	*0.72*	*0.87*
MATH	*H*	*150+*	*530*	*5.05*	*3.47*	*2.85*	*0.77*	*0.78*	*0.85*

Note. MAD=mean absolute difference. EBLP=empirical best linear predictor. Sample includes all school by disaggregated group combinations with direct and EBLP estimates in both 2024 and 2025. E=elementary (grades 4-5); M=middle school (grades 6-8); H=high school (grades 9-11, excluding grade 9 ELA).

To gain additional insight into how the EBLP method functions, Figure 1 plots the difference between direct mean SGP and EBLP estimates across schools when aggregating SGP for all students at each school. The figure combines math and ELA results and plots results separately by EMH level, so that each marker represents the difference between the direct observed mean SGP and the EBLP mean SGP for all students at a school in either math or ELA. The figure shows that differences between direct and stabilized estimates are largest for the schools with the fewest students, and become systematically smaller as enrollment size increases. This figure thus provides some intuition about why the EBLP method increases stability more for smaller schools.

Figure 1: Difference between school direct mean SGP and EBLP SGP estimates in 2025 versus number of students with SGP



Note. E=elementary (grades 4-5); M=middle school (grades 6-8); H=high school (grades 9-11, excluding grade 9 ELA). Each point represents the All Student Group for a single school in either math or ELA. All schools with direct and EBLP estimates in both 2024 and 2025 are represented.

Accuracy of Direct versus EBLP Stabilized Estimates

Table 7 reports the relative accuracy of direct mean scale scores and mean SGP relative to the EBLP mean scale scores and SGP. The results are again presented for the All Student groups sample, where each unit is a school by disaggregated group. Accuracy ratios greater than 1.0 indicate that the EBLP mean estimate is a more accurate estimate relative to the direct mean estimate. As expected, the EBLP mean scale score and mean SGP are more accurate than the direct mean estimates across all subjects, EMH levels, and school sizes. The accuracy of the EBLP estimates was higher than for the direct estimate for all schools in the sample. The average accuracy ratios range from 1.10 to 1.61 (with an average of 1.34) for mean scale scores and range from 1.06 to 1.72 (with an average of 1.36) for mean SGP. Consistent with the EBLP approach having a larger impact on smaller schools, the mean accuracy ratios were higher for the smaller sample size groups.

Table 7: Accuracy ratio of school mean scale scores and mean SGP relative to stabilized EBLP estimates

		Scale Scores			SGP		
Subject	**EMH**	**Size**	**Schools**	**Accuracy Ratio**	**Size**	**Schools**	**Accuracy Ratio**
ELA	*E*	*16-49*	*1423*	*1.45*	*20-49*	*1023*	*1.47*
ELA	*E*	*50-149*	*1582*	*1.26*	*50-149*	*1411*	*1.24*
ELA	*E*	*150+*	*673*	*1.15*	*150+*	*135*	*1.11*
ELA	*M*	*16-49*	*554*	*1.48*	*20-49*	*465*	*1.56*
ELA	*M*	*50-149*	*671*	*1.30*	*50-149*	*646*	*1.28*
ELA	*M*	*150+*	*688*	*1.14*	*150+*	*622*	*1.10*
ELA	*H*	*16-49*	*402*	*1.57*	*20-49*	*297*	*1.72*
ELA	*H*	*50-149*	*372*	*1.44*	*50-149*	*338*	*1.65*
ELA	*H*	*150+*	*434*	*1.29*	*150+*	*377*	*1.40*
MATH	*E*	*16-49*	*1424*	*1.35*	*20-49*	*979*	*1.34*
MATH	*E*	*50-149*	*1560*	*1.19*	*50-149*	*1503*	*1.14*
MATH	*E*	*150+*	*702*	*1.10*	*150+*	*147*	*1.06*
MATH	*M*	*16-49*	*547*	*1.53*	*20-49*	*465*	*1.60*
MATH	*M*	*50-149*	*664*	*1.32*	*50-149*	*644*	*1.32*
MATH	*M*	*150+*	*704*	*1.16*	*150+*	*627*	*1.12*
MATH	*H*	*16-49*	*410*	*1.61*	*20-49*	*309*	*1.66*
MATH	*H*	*50-149*	*370*	*1.47*	*50-149*	*368*	*1.43*
MATH	*H*	*150+*	*442*	*1.33*	*150+*	*530*	*1.18*

Note. Sample includes all school by disaggregated group combinations with direct and EBLP estimates in both 2024 and 2025. E=elementary (grades 3-5); M=middle school (grades 6-8); H=high school (grades 9-10 for scale scores and grades 9-11, excluding grade 9 ELA, for SGP).

Comparison of SPF Ratings

This final analysis provides a preliminary exploration of the impact that using stabilized EBLP estimates to summarize SGP might have on school SPF growth ratings. The results above demonstrate that, as expected, the EBLP mean SGP estimates are more stable across years relative to direct mean SGP and median SGP. However, school SPF growth ratings are calculated through a multi-step process that combines information from aggregate SGP of multiple student groups. These aggregate SGP for each group are scored according to a rubric, weighted, and then combined into an overall rating. The SPF growth rating is then combined with a similar rating for achievement (and a rating for PWR for high schools). It is not immediately clear how use of the stabilized EBLP estimates would impact the year-to-year variability of the scored SPF ratings due to the multi-step nature of this process.

To provide a preliminary exploration of this question, we compared how school SPF growth ratings varied from 2024 to 2025 when calculated in three different ways:[11]

[11] The results described here were prepared by CDE. Results in Tables 8 and 9 are adapted from results presented to the Colorado Technical Advisory Panel at the October 2025 meeting. Materials available here: https://ed.cde.state.co.us/accountability/stakeholders/tap

1. Original ratings using direct median SGP, with original scoring rules.
2. Ratings using direct mean SGP estimates, with modified scoring rules.
3. Ratings using EBLP mean SGP estimates, with modified scoring rules.

The original scoring rules for the SPF assign points based on where the observed median SGP for a particular group is relative to threshold values of 35, 50, and 65.[12] These threshold values used to assign SPF ratings (35, 50, and 65) correspond roughly to the 15th, 50th, and 85th percentile of the distribution of median SGP across schools. Because the distribution of school mean SGP differs from the distribution of school median SGP, the threshold values of 35/50/65 would not be appropriate for use with mean SGP. As shown in Table 4 above, the distribution of mean SGP across schools is narrower than the distribution of median SGP. As a result, while the 15th percentile of school median SGP is approximately 35, the 15th percentile of school mean SGP is approximately 40. Thus, to create school SPF growth ratings based on mean SGP that would be comparable to SPF growth ratings based on median SGP, we used threshold values of 40/50/60, which correspond roughly to the 15th, 50th, and 85th percentiles of the school mean SGP distribution, to assign ratings. We refer to using these alternative threshold values as the modified scoring rules. We used the same modified scoring rules for direct mean SGP and EBLP mean SGP. The analysis is based on a sample of 1,904 schools with sufficient reportable data and combines across all EMH levels.

We emphasize that this is only a preliminary analysis. Transitioning from the use of median SGP to direct mean or EBLP mean SGP represents a fundamental shift to the SPF methodology and scoring rules that would require additional potential changes. This analysis was carried out to provide intuition of potential impacts. These results are also based on preliminary rather than final SPF ratings for the associated years.

Table 8 summarizes the result of this experiment. The table reports how each school's SPF growth rating would change from 2024 to 2025 with one of the three different calculation methodologies listed above (original median ratings, direct mean ratings, or EBLP mean ratings). Based on the original ratings, 74.5% of schools retained the same rating from 2024 to 2025, 14.2% of schools saw an increase in their rating, and 11.4% saw a decrease. These values were 71%, 15.8%, and 13.2%, respectively, when using the direct mean SGP. Thus, although direct mean SGP were more stable across years than the direct median SGP, SPF growth ratings based on the direct mean SGP were less stable across the two years. Finally, when using the EBLP mean SGP, 74.4%, 13.2%, and 11.5% of schools saw no change, increases, or decreases in SPF growth ratings from 2024 to 2025. These values are very similar to the values based on direct median SGP. Table 9 reports the percent of schools that would maintain the same SPF growth rating across 2024 to 2025 using the three different scoring approaches, separately by school size and EMH level. The table shows that while smaller schools tend to have more variable SPF ratings across years, the pattern of results across the three methods is the same for schools of different sizes or EMH levels. High school SPF growth ratings are considerably more stable across all three methods compared to elementary and middle schools. This is likely due to a combination of factors including higher average sample sizes and differences in the distribution of mean and median SGP relative to the scoring thresholds at the high school level.

This preliminary analysis suggests that although the mean SGP and EBLP stabilized SGP metrics are more stable from year to year than the median SGP, use of these alternative SGP measures may not necessarily lead to greater stability of overall SPF ratings. While this result is somewhat counterintuitive, there are likely two primary factors that lead to this result. First, the threshold values used to score the observed school median or mean SGP are normative (based on the

[12] See the SPF Scoring Guide for more details: SPF Scoring Guide: https://www.cde.state.co.us/accountability/frameworkscoringguide

15th, 50th, and 85th percentile of the across school distributions). Thus, although mean SGP are less variable across years, the modified scoring thresholds partly take this into account. Second, school SPF ratings combine results from multiple indicators across multiple groups and use a multi-step process to produce an overall rating. These results emphasize that even when each individual indicator is more stable across years, the combination of these indicators constructed in the SPF process may not necessarily be more stable across years.

Table 8: Comparison of SPF Growth Ratings using median, mean, or EBLP stabilized SGP

		2025 Rating		
		Using Original Scores (Median SGP w/ no Stabilization) Rating Change: 74.5% Retain, 14.2% Increase, 11.4% Decrease		
2024 Rating	**Performance**	**Improvement**	**Priority Improvement**	**Turnaround**
Performance	*1189*	*120*	*16*	*2*
Improvement	*143*	*171*	*48*	*15*
Priority Improvement	*29*	*57*	*43*	*16*
Turnaround	*5*	*18*	*17*	*15*

		2025 Rating		
		Using Direct Scores (Mean SGP w/ no Stabilization) Rating Change: 71% Retain, 15.8% Increase, 13.2% Decrease		
2024 Rating	**Performance**	**Improvement**	**Priority Improvement**	**Turnaround**
Performance	*1121*	*134*	*17*	*7*
Improvement	*167*	*164*	*59*	*14*
Priority Improvement	*21*	*66*	*47*	*21*
Turnaround	*4*	*17*	*25*	*20*

		2025 Rating		
		Using Stabilized Scores (Mean SGP w/ EBLP stabilization) Rating Change: 74.4% Retain, 13.2% Increase, 11.5% Decrease		
2024 Rating	**Performance**	**Improvement**	**Priority Improvement**	**Turnaround**
Performance	*1151*	*117*	*13*	*1*
Improvement	*151*	*193*	*50*	*13*
Priority Improvement	*13*	*63*	*50*	*23*
Turnaround	*1*	*12*	*30*	*23*

Note. Total sample of 1,904 schools. Ratings for direct mean SGP and EBLP SGP use modified SPF scoring rules described in the main text. All results are based on preliminary SPF data.

Table 9: Stability of SPF growth ratings using median, mean or EBLP stabilized SGP by school size and level

School Category	Median SGP	Mean SGP	EBLP SGP
All Schools	*74.5%*	*71.0%*	*74.4%*
Large (N>=400)	*80.5%*	*77.5%*	*79.1%*
Medium (150<=N<400)	*70.3%*	*67.1%*	*71.8%*
Small (N<150)	*72.6%*	*67.2%*	*71.4%*
Elementary	*70.8%*	*66.6%*	*71.3%*
Middle	*72.4%*	*70.6%*	*74.6%*
High School	*87.8%*	*83.9%*	*82.8%*

Note. Stability is measured as the percent of schools with the same SPF growth rating in 2024 and 2025 using different growth metrics. Data are for the same sample of 1,904 schools reported in Table 8.

Discussion

Returning to the framework provided by the Colorado 1241 Task Force, the analyses presented here relate directly to the matter of volatility of test score metrics in small groups. The empirical results demonstrate that EBLP stabilization meaningfully reduces year-to-year variability in school test score metrics, with the largest gains observed for schools and subgroups with the smallest sample sizes, which are the populations highlighted in Recommendation #4. Specifically, applying the EBLP approach increased year-to-year stability of mean scale scores and mean SGP, as reflected in higher correlations across years and lower mean absolute differences (MAD). These patterns align with the intended features of the EBLP method – reducing variability in estimates when sample sizes are small – while still allowing for meaningful differences across units. We observed some variation in the magnitude of impacts by subject and EMH level, which may be due to differences in school sample size distributions and test characteristics and could be an avenue for further investigation. Although we could not apply the EBLP stabilization to median SGP, we found that the direct mean SGP tended to be more stable than median SGP, consistent with common statistical expectations. Some state guidance explicitly cites this reliability advantage when choosing mean SGP over median SGP for higher-stakes uses. For example, the Georgia Department of Education notes that “means are more statistically reliable than medians,” particularly for smaller groups (Georgia Department of Education, 2023). At the same time, other states continue to use median SGP to preserve continuity and interpretability for stakeholders.

When we examined implications of using EBLP stabilized metrics for SPF growth ratings, substituting mean SGP for median SGP increased year-to-year volatility, while using an EBLP mean reduced volatility relative to the direct mean, but yielded volatility similar to use of median SGP. Although potentially counterintuitive, this result underscores an important practical point: because SPF ratings are produced via multi-step combinations of indicators, improving the year-to-year stability of the individual indicators does not guarantee improved stability of overall SPF ratings. This analysis should be considered preliminary, because changing growth metrics would require additional policy decisions and technical updates across the SPF system and could change the conclusions.

We end by noting some additional points relevant to considering the use of EBLP stabilized test score metrics in the context of Colorado's school accountability system and limitations of the current analysis. First, the statistical methodology underlying EBLP is premised on estimating an underlying "true" quantity (e.g., a school or district's population mean test score or mean SGP), treating observed annual summaries as imperfect estimates that include sampling and measurement error. This logic is consistent with established psychometric frameworks that model uncertainty in aggregate educational indicators and emphasize that observed summaries can vary meaningfully due to measurement and sampling variability rather than true change (Cronbach et al., 1997). However, this framing may differ from how accountability indicators are interpreted in practice, where stakeholders may view each year's reported result as the outcome to be judged rather than a noisy estimate of a stable underlying parameter.

Second, while a theoretical benefit of EBLP is improved stability, which may support decision making that depends on distinguishing sustained performance patterns from idiosyncratic year-to-year fluctuation, a key tradeoff is transparency and explainability. The EBLP estimates are less straightforward to communicate because they incorporate information beyond the focal year via a model fit to statewide data. In addition, direct (non-EBLP) summaries are easier for districts and schools to reproduce locally from student-level records, whereas EBLP estimates cannot be reconstructed without the statewide model and data.

Third, our analysis implemented only one statistical stabilization approach, EBLP, because we considered it the most appropriate method for this context. Bayesian stabilization has been proposed as a method to improve reliability of test score metrics for small samples (Forrow et al., 2023; Rosendahl et al., 2024). However, we considered Bayesian stabilization as implemented in these prior studies less appropriate than EBLP because it uses only aggregate rather than student-level data, applies adjustment primarily toward the state average (raising fairness concerns relative to EBLP's adjustment towards school-specific data), and had not been operationally adopted by any state nor supported by publicly available software at the time of our analyses.[13] Some states apply shrinkage within value-added models to reduce year-to-year variation in school growth scores (e.g., Wisconsin; Education Analytics, 2025). While this type of shrinkage approach shares similarities with the EBLP method, this approach could not be applied directly to Colorado data given that Colorado does not use a value-added model. Future work investigating the use of alternative stabilization approaches would be useful.

Fourth, two limitations may affect the generalizability of our findings even within the Colorado context. First, the analysis examined SPF growth ratings only, not overall SPF ratings. Statistical stabilization may have larger or smaller effects on overall SPF ratings, which combine multiple indicators. Second, our analyses used data from 2022–23, 2023–24, and 2024–25 because these were the most recent data available. Test scores in these years were likely still shaped by COVID-19 disruptions to testing and student learning. For example, skip-grade testing in 2021 reduced the number of prior scores available for 2023 SGP calculations and test participation rates, while recovering, remained below pre-2020 levels in some cases. Additionally, year-to-year changes in student achievement during this period may differ from what is expected in future years. As pandemic-related effects recede and achievement patterns settle, the impacts of stabilization may differ in future years.

[13] An R package to implement the Bayesian stabilization method was subsequently released in March 2026. However, consistent with the prior reports, this package uses only aggregate data. The package can be accessed at: https://ies.ed.gov/use-work/resource-library/resource/other-resource/accuracy-everyone-r-package.

Finally, statistical stabilization is only one possible avenue for addressing volatility. Other options include increasing minimum n-size thresholds for inclusion or changing SPF calculation methods (e.g., multi-year averaging or modifying the included indicators and scoring rules). These have not currently been proposed (in fact there has been consideration of reducing, rather than increasing, the minimum n-size) and would entail more substantial changes to Colorado's school accountability system.

References

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

Education Analytics. (2025). *Technical report on the Wisconsin value-added model: Academic year 2024-25.* https://dpi.wi.gov/sites/default/files/imce/accountability/pdf/WI_DPI_School_VA_Technical_Report_2025_Final.pdf

Forrow, L., Starling, J., & Gill, B. (2023). *Stabilizing subgroup proficiency results to improve the identification of low-performing schools* (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

Georgia Department of Education. (2023). *Methods of combining SGPs*. https://lor2.gadoe.org/gadoe/file/1515feae-3c09-4d73-8976-82698f17e86f/1/Methods_of_Combining_SGPs.pdf?utm_source=chatgpt.com

Lockwood, J. R., & Castellano, K. E. (2022). *schoolgrowth: Functions for improving accuracy of school-level growth measures using Empirical Best Linear Prediction* (Version R package version 0.6-3) [Computer software]. https://github.com/EducationalTestingService/schoolgrowth

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

McCaffrey, D. F., & Castellano, K. E. (2021, December 8). *A new approach for reporting aggregate student growth scores*. https://www.ets.org/news/stories/a-new-approach-for-reporting-aggregate-student-growth-scores.html

R Core Team. (2025). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rosendahl, M., Gill, B., & Starling, J. E. (2024). *Stabilizing school performance indicators in New Jersey to reduce the effect of random error* (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