COLORADO
Department of Education

Technical Advisory Panel Meeting

September 23, 2025

Welcome & Introductions

- **Welcome from CDE**
  - The purpose of the TAP is to provide non-binding technical recommendations to CDE regarding the Colorado Growth Model, state accountability, and other topics as needed.
- **Meeting Logistics:**
  - Non-members, please add your Name/Affiliation to the chat box.
  - Everyone please mute your sound.
  - We ask all non-TAP members to hold any comments until the end of the meeting. We do this to ensure we have sufficient time to address all meeting agenda items.
- **Introductions with Scott Weldon, TAP Chair**
  - Welcome Jessica Alzen!

Agenda for Today

- **Welcome and Introductions** | Information Item
- **2026 Performance Framework Targets** | Feedback Item
- **Recap and Roadmap for HB25-1278** | Feedback Item
- **Small N Stabilization Study Background** | Feedback Item
- **Wrap-Up**

2026 Performance Framework Targets

Marie Huchton

Feedback Item

Considerations for 2026 Target Setting

Substantial increases in 2025 SAT mean scale scores, call into question what Achievement targets (i.e. sub-indicator cut-scores) should be used for 2026.

In theory, now that we have three years of digital PSAT/SAT data, the 1 and 3-year results should be similar enough to use the same targets

- However, early years of a new assessment often see an implementation dip, so we can't be sure the 3-year results won't be slightly lower (at least for PSAT)
- Given the score fluctuations we saw in 2025, no way to predict what will happen for 2026

Differences in 2024 and 2025 PSAT/SAT Scale Scores

			Calculated Scale Score Cut-Points		Difference	
			2024 1yr	2025 1 yr	2024 to 2025	2026 1 & 3yr
Achievement: PSAT & CoAlt DLM Grades 9 & 10	Reading & Writing	15th	415.1	414.4	-0.7	
		50th	458.9	454.9	-4.0	
		85th	505.0	508.4	3.4	
	Math	15	387.4	387.8	0.4	
		50	430.2	431.0	0.8	
		85	480.4	482.4	2.0	
PWR: SAT & CoAlt DLM Grade 11	Reading & Writing	15	448.1	456.9	8.8	
		50	494.6	504.3	9.7	
		85	553.1	558.0	4.9	
	Math	15	423.3	423.2	-0.1	
		50	465.8	470.0	4.2	
		85	527.0	531.7	4.7	

?

How do we want to approach 2026 Target setting?

Options for 2026 Target Setting

Option 1: Hold- Maintain existing targets, and communicate to districts that no matter what the 2026 results look like (better or worse), we will not be changing the targets/subsequent ratings.

Option 2: Increase & revise- Update SAT targets to match the higher 2025 distribution (best guess on what future years will look like). Once 2026 results become available, revisit any of the revised targets that are out of alignment with 2026 percentile rank results.

Option 3: Wait & revise- Leave 2026 SAT Targets as TBD (like we did the 3-year targets last year) and communicate that once 2026 results become available, percentile rank targets will be established at the 15th-50th-85th percentiles by test and content area

TAP RECOMMENDATION

Given the limited information currently available, what approach does the TAP recommend for proposing 2026 achievement targets to the State Board of Education?

HB25-1278 Education Accountability System

Lisa Medler

Feedback Item

Background | HB 25-1278: Education Accountability System

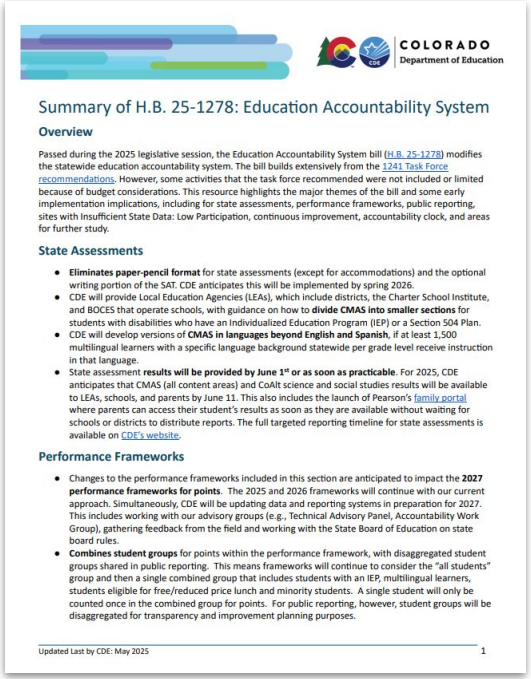
Policy Development Structures
State Legislature
State Board of Education
CDE Documentation and Guidance

- Builds from the 1241 Task Force Recommendation Report. All recommendations were generally included in the final version of the bill, with the exception of some of the studies (e.g., new rating labels, accreditation system, awards, best practices).
- (May 23) Signed bill
- Several reports to policymakers
- Implications for State Board rules and CDE guidance

Background | HB 25-1278 Resources

HB 25-1278 Fact Sheet - Provides a high level overview of the key themes of the accountability bill, which include:

- State Assessments
- Performance Frameworks
- Postsecondary & Workforce Readiness Indicator
- Public Reporting
- Insufficient State Data: Low Participation
- Continuous Improvement (including the Accountability Clock and various supports)
- Stakeholder Groups & Studies



COLORADO
Department of Education

Summary of H.B. 25-1278: Education Accountability System

Overview

Passed during the 2025 legislative session, the Education Accountability System bill (H.B. 25-1278) modifies the statewide education accountability system. The bill builds extensively from the 1241 Task Force recommendations. However, some activities that the task force recommended were not included or limited because of budget considerations. This resource highlights the major themes of the bill and some early implementation implications, including for state assessments, performance frameworks, public reporting, sites with Insufficient State Data: Low Participation, continuous improvement, accountability clock, and areas for further study.

State Assessments

- **Eliminates paper-pencil format** for state assessments (except for accommodations) and the optional writing portion of the SAT. CDE anticipates this will be implemented by spring 2026.
- CDE will provide Local Education Agencies (LEAs), which include districts, the Charter School Institute, and BOCES that operate schools, with guidance on how to **divide CMAS into smaller sections** for students with disabilities who have an Individualized Education Program (IEP) or a Section 504 Plan.
- CDE will develop versions of **CMAS in languages beyond English and Spanish**, if at least 1,500 multilingual learners with a specific language background statewide per grade level receive instruction in that language.
- State assessment **results will be provided by June 1st or as soon as practicable**. For 2025, CDE anticipates that CMAS (all content areas) and CoAlt science and social studies results will be available to LEAs, schools, and parents by June 11. This also includes the launch of Pearson's family portal where parents can access their student's results as soon as they are available without waiting for schools or districts to distribute reports. The full targeted reporting timeline for state assessments is available on CDE's website.

Performance Frameworks

- Changes to the performance frameworks included in this section are anticipated to impact the **2027 performance frameworks for points**. The 2025 and 2026 frameworks will continue with our current approach. Simultaneously, CDE will be updating data and reporting systems in preparation for 2027. This includes working with our advisory groups (e.g., Technical Advisory Panel, Accountability Work Group), gathering feedback from the field and working with the State Board of Education on state board rules.
- **Combines student groups** for points within the performance framework, with disaggregated student groups shared in public reporting. This means frameworks will continue to consider the "all students" group and then a single combined group that includes students with an IEP, multilingual learners, students eligible for free/reduced price lunch and minority students. A single student will only be counted once in the combined group for points. For public reporting, however, student groups will be disaggregated for transparency and improvement planning purposes.

Updated Last by CDE: May 2025 1

Guiding Principles for Implementation (from State Board of Education at the December 2024 meeting on priorities for HB25-1278)

COHERENCE	TRANSPARENCY	CONSISTENCY	SUPPORT	FLEXIBILITY
Ensure **coherence across changes to the system** by building upon areas of strength from the current system to maintain high expectations for all Colorado students. Changes should be intentionally sequenced, rather than taking a fragmented or a piecemealed approach. Align, where possible, with other task force recommendations (e.g., 1215 Task Force) and federal accountability expectations. Take efforts to balance the cost of implementing change with the potential long-term impacts.	Enhance **transparency and trust of the accountability system**, including access to information for families, schools, and the general public.	Ensure **statewide consistency** using multiple measures that meaningfully differentiate sites to guide resources and supports.	Strengthen the continuous improvement approach for all sites by ensuring a **continuum of resources and supports**. This includes recognizing performance and identifying bright spots to foster shared learning, proactively supporting schools to help them avoid entering the accountability clock, and expanding end-of-clock options.	Ensure that **policies are flexible enough so that adjustments can be made over time**, based on stakeholder feedback and ongoing research.

Implementation Phases of the Accountability Bill

Phase 1: Discovery and Preparation

2025-26

Focus on studies, stakeholder engagement, and design work.
Note: Some components begin sooner.

Phase 2: Pilot and Communication

2026-27

Drive new initiatives and maintain open communication.

Phase 3: Launch Full Implementation

2027-28

Fully implement changes and establish ongoing continuous improvement cycles.

Timeline | What will 2025-26 look like?

Phase 1:
Discovery and Preparation
2025-26

State Board

- Accountability Hearings for End of Clock sites
- **Spring 2026**: Rulemaking ISD, frameworks, and end-of-clock
- Initial studies

CDE

- Studies: small system, assessments, dashboard
- Launch refreshed AWG and create opportunities for stakeholder engagement
- Begin studying framework changes

Districts/LEA

- Transition away from paper-based assessments (except for students with accommodations in their IEPs) for all LEAs
- Participate in feedback opportunities

Timeline | What will 2026-27 look like?

Phase 2:
Pilot and Communication

2026-27

State Board

- **Fall 2026:** Rulemaking on PWR & frameworks
- **Winter 2026:** Incorporate recommended changes into 2027 target setting

CDE

- **Fall 2026:** Small systems, Assessment and Dashboard Studies complete
- Recommendations to SBE on incorporating changes

Districts/LEA

- **Fall 2026:** Insufficient State Data: Low Participation Year 1 begins
- **Spring 2027:** Informational frameworks

Timeline | What will 2027-28 look like?

Phase 3:
Launch Full Implementation

State Board

- Full implementation (e.g., target setting, finalize frameworks)

CDE

- **Fall 2027:** Evaluation of external managers
- Support for Insufficient State Data presentations
- Ongoing technical assistance and support

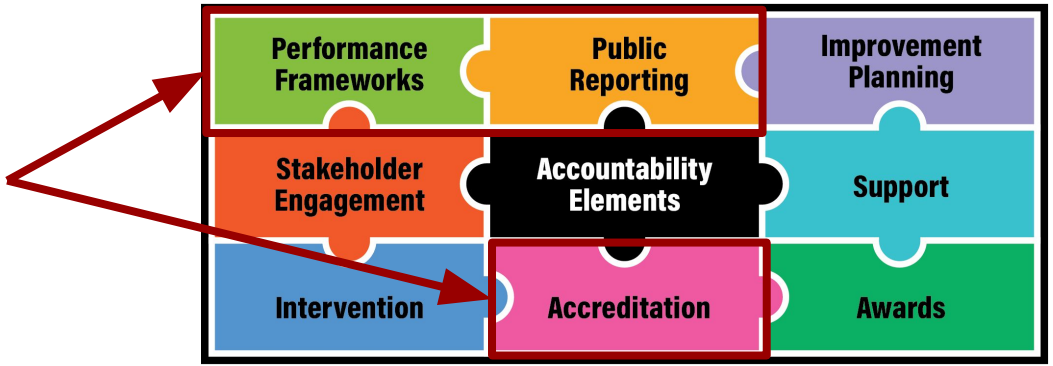
Districts/LEA

- **Fall 2027:** Framework changes implemented
- **Spring 2028:** Begin preparation for presentations to the State Board of Education for Insufficient State Data: Low Participation.

What is TAP's role?

Since the passage of the Education Accountability Act of 2009, the TAP has held a crucial role as **technical experts** for the department. HB 25-1278 mentions TAP several times as a key stakeholder group to support implementation. All TAP recommendations are **non-binding** (i.e., we cannot guarantee that TAP recommendations will result in changes).

Typically, TAP recommendations address these three elements.



TAP 1278 Conversation Topics

Continued Topics
New Topic
Recommendation Needed

Meeting	Tentative Agenda Topics	Notes
Sept 23	1278 Implementation 2026 Framework Targets Small System Studies	● We last discussed HB 25-1278 and small system studies during the May 2025 meeting
Oct 23	Small System Studies On Track Growth for HS and 9th grade ELA Analysis for Former IEP Students	
Nov 20	**Small System Studies** **On Track Growth** Analysis for Former IEP Students ISD: Low Part & Assessment Update	● To run calculations for the next agenda topics, a recommendation is needed for small systems and On Track Growth
TBD Dec	**Analysis for Former IEP Students** ISD: Low Part & Assessment Update **Dashboard Study**	● We likely need to schedule a December TAP meeting ● To run calculations about a Combined Student Group, a recommendation is needed for Former IEP

TAP 1278 Conversation Topics

Continued Topics
New Topic
Recommendation Needed

Meeting	Tentative Agenda Topics	Notes
Jan 20	**ISD: Low Part & Assessment Update** PWR Indicator Combined Student Group	● If needed, a follow-up about OTG may delay other topics ● To inform State Board rules in the spring, a recommendation is needed for ISD: Low Participation
Feb 24	PWR Indicator Combined Student Group	
Mar 24	PWR Indicator **Combined Student Group**	● To inform State Board rules in the fall, a recommendation is needed
Apr 28	PWR Indicator Framework Weights Request to Reconsider for Distinction	
May 21	**PWR Indicator** Framework Weights **Request to Reconsider for Distinction**	● To inform State Board rules in the fall, a recommendation is needed

TAP RECOMMENDATION

Given these topics, what questions do you have for CDE? What else should we consider?

Use this link so we can schedule a December TAP meeting **(or use the QR code)**

Small N Stabilization Study Background

Daniel Mangan & Ben Shear

Feedback Item

Overview

Background	• Legislative Background • Importance of Min N Size • National/Colorado context
Review Reliability Study	• Research Questions & Findings • See resources from May 2025 TAP meeting for more info
Introduce Stabilization Study	• Stabilization Goals and Context • EBLP Methodology • Research Questions & Findings
Discussion	• What are you taking away from these two studies? • Questions/concerns for follow up?

Legislative Background

1241: Accountability, Accreditation, Student Performance, and Resource Inequity Task Force Report

1241 Task Force Recommendations

#1: Lower student count thresholds for accountability calculations and reporting →

#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 →

CRS 22-11-212(1)(a) and (2)(a)

(1)(a) THE DEPARTMENT SHALL FACILITATE A STUDY, IN CONSULTATION WITH THE TECHNICAL ADVISORY PANEL, A COUNCIL THAT FOCUSES ON RURAL EDUCATION, THE ACCOUNTABILITY WORK GROUP DESCRIBED IN SECTION 22-11-202 (3), AND OTHER ADVISORY GROUPS WITH RELEVANT EXPERTISE, AND **MAKE RECOMMENDATIONS ON LOWERING STUDENT COUNT THRESHOLDS ON ACCOUNTABILITY CALCULATIONS AND REPORTING**....

(2)(a)...**MAKE RECOMMENDATIONS ON ADDRESSING INHERENT VOLATILITY OF TEST SCORE MEASUREMENTS FOR LOCAL EDUCATION PROVIDERS WITH SMALL STUDENT POPULATIONS**.

Why Minimum Group Size Matters

Transparency & Equity

- Lower thresholds allow more reporting so progress for small student groups is more visible.

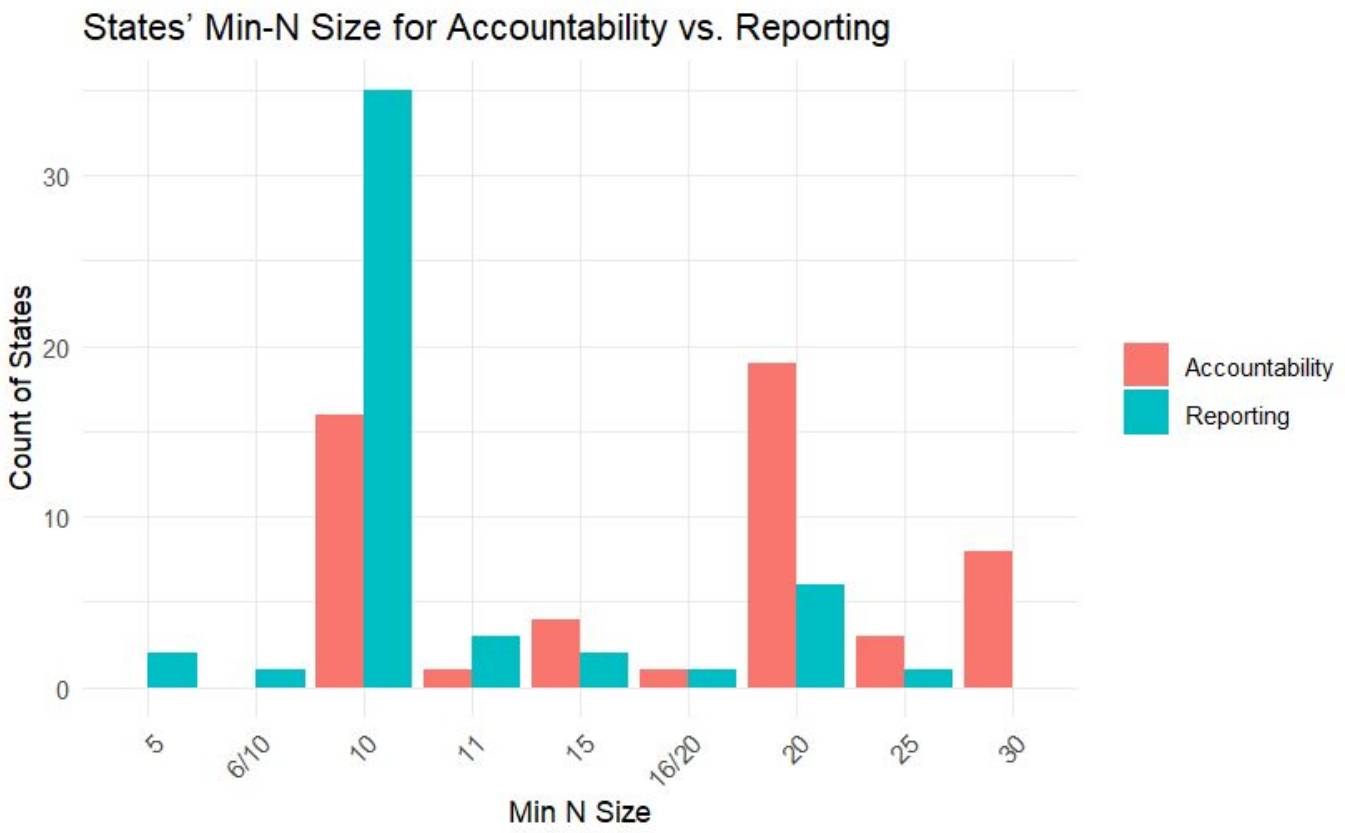
Fairness & Privacy

- Small student groups increase volatility in scores, meaning they are less reliable representations of true performance.
- Risks identifying individual students.

National Context: No Single "Right" Answer Regarding Minimum N Size

So, a state should pick a fairly small *N* for purposes of validity (say, certainly something no larger than 10), but it would need a very high *N* for purposes of reliability (say, 300 or more). Obviously, a value that provides reasonable validity is wholly inadequate for reliability purposes; a value that provides reasonable reliability is wholly inadequate for validity purposes. A figure between those two is largely inadequate for *both* purposes. This is the reason states are having such a hard time choosing a fixed value for minimum *N*. At first blush, it appears that the problem is choosing a modest fixed *N* that is a reasonable compromise between reliability and validity; a careful look tells us that choosing any value is wholly inadequate for at least one of the two concerns, if not both. In short, there is not a reasonable answer to this dilemma. One is not faced with a reasonable balancing of concerns over reliability and validity; any answer will be clearly wrong for at least one of the two.

Hill & DePascale, 2003



Alliance for Excellent Education, 2018

Colorado Context

- Colorado currently requires at least 16 students for achievement scores and 20 for growth scores to count in school ratings.
- These requirements have been in place since 2010.
- Two primary concerns w/ current requirements:
  - When minimum n-sizes are not met, it is sometimes not possible to construct accountability ratings (ISD ratings) and/or disaggregated group results don't get reported.
    - This can undermine transparency/equity goals that rely on reporting data for all student groups
  - When student counts are at or just above minimum n-sizes, aggregate test score results can be highly variable from year to year.
    - This can be perceived as unfair, since ratings can be overly influenced by a small number of students and/or factors outside the control of schools/districts.

(Shear et al., 2025)

Overview

Background	• Legislative Background • Importance of Min N Size • National/Colorado context
Review Reliability Study	• Research Questions & Findings • See resources from May 2025 TAP meeting for more info
Introduce Stabilization Study	• Stabilization Goals and Context • EBLP Methodology • Research Questions & Findings
Discussion	• What are you taking away from these two studies? • Questions/concerns for follow up?

Reliability Study Research Questions

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?

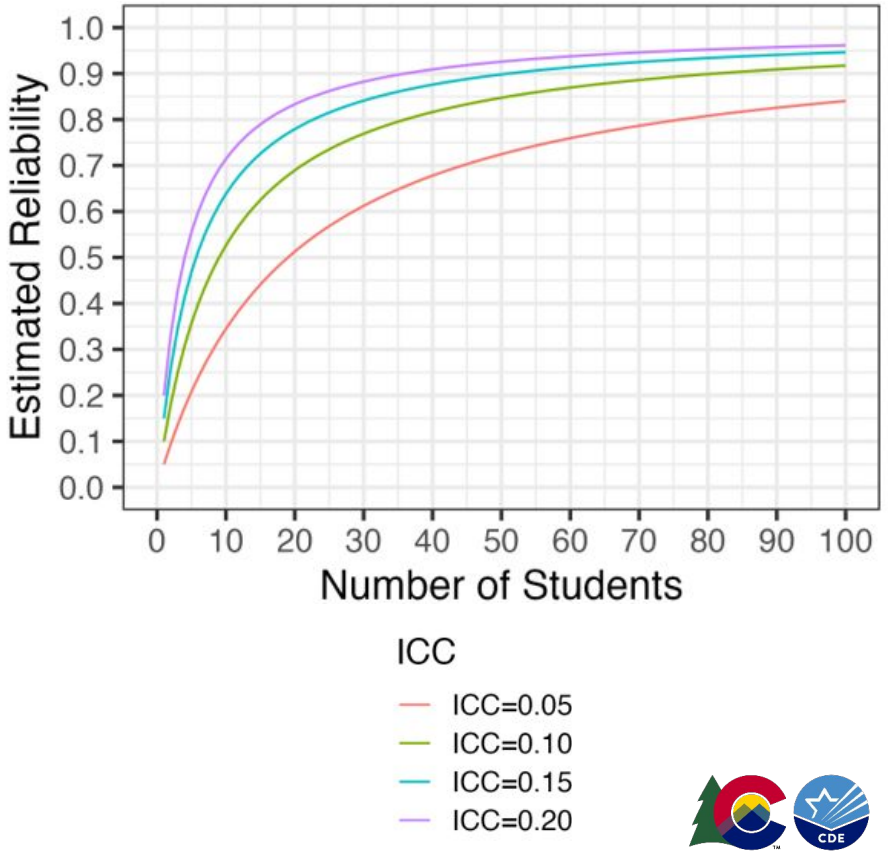
Analytic Approach: We fit a two-level hierarchical linear model of student test scores nested within schools to estimate between- and within-school variance, compute intraclass correlations, and derive school-level reliability coefficients adjusted for each school's sample size.

RQ1 Results: Sample Size Impacts

- Reducing min-n threshold to 10 for **achievement** would increase reportable schools by 2-3 percentage points and districts by 4-5 percentage points.
- Reducing min-n threshold to 10 for **growth** would increase reportable schools by 2-5 percentage points and districts by 6-7 percentage points .
- For disaggregated groups, range of sample size impact is much larger due to variability in group size
  - Achievement: 3 to 15 pct point increase in share of schools w/ reportable data schools.
  - Growth: 6 to 45 pct point increase in share of schools w/ reportable data schools.
- Largest increases in reportable data for ML and IEP disaggregated groups

RQ2 Results: ICC & Reliability

- **Reliability**: the share of a measure that reflects true differences (signal) rather than chance fluctuations (noise)
- Ranges from 0 – 1; 0.8 is often considered “good” but this is a highly context-driven heuristic
- Reliability is driven by two main factors:
  - **Intraclass Correlation** (ICC): how much signal is coming from the school level?
  - Sample Size
    - Larger n noise cancels out, averages stabilize
    - Smaller n a few extreme scores can swing the mean



Reliability Study Takeaways

- Although lowering minimum n-size thresholds to 10 would allow for increases of about 2-7 pct points in share of schools/districts with reportable data, measures for smaller groups/systems would be less reliable.
- Whether increased reporting would be worth reduced reliability requires considering how results would be used to improve opportunities for students enrolled in these schools/districts.
- Reliability of mean scale scores is higher than for average SGP for equivalent n sizes, which supports having higher minimum-n for growth.
- School metrics tend to be more reliable than district metrics for same sample size, which could suggest setting different thresholds for schools and districts.
- Reliabilities calculations are a lower bound – most units will have larger n-sizes and thus higher reliability. There will also be variation from year to year, making it important to choose a proper baseline and make decisions about whether same thresholds are used across units and/or are allowed to vary.

Overview

Background	• Legislative Background • Importance of Min N Size • National/Colorado context
Review Reliability Study	• Research Questions & Findings • See resources from May 2025 TAP meeting for more info
Introduce Stabilization Study	• Stabilization Goals and Context • EBLP Methodology • Research Questions & Findings
Discussion	• What are you taking away from these two studies? • Questions/concerns for follow up?

Stabilization Study: The Issue

- Variability of aggregate test score metrics (median SGP, average scale scores) for small groups, schools, and districts. Larger concern for growth metrics than average test scores.
- **HB25-1278:** Requires the department, in consultation with the technical advisory panel, the accountability work group, and other advisory groups with relevant expertise, **to study** lowering student count thresholds on accountability calculations and reporting, **addressing inherent volatility of test score measurements for local education providers with small student populations**…

The Goal of “Stabilization” Approaches

- Reduce the year-to-year volatility of test score metrics particularly for small groups.
- Create more accurate estimates of the “true” long-term quantity of interest (e.g., median SGP [MGP] or mean scale score).
- The intention of these methods is to use an “infinite population” statistical framework to produce more accurate and stable estimates for small groups.

Conceptual Mechanics of Stabilized Estimates

- Stabilization works by utilizing additional data (or information) beyond the test scores or SGP observed for students in the current year.
- There are two primary sources of additional information:
  - Data from **other similar schools/students in the same year**.
  - Data from students in the **same school in prior years**.
- CDE believes it is more appropriate to incorporate data from the same school in prior years.
  - Multi-year framework ratings can be seen as a form of stabilization (essentially, a sample-size weighted average over three years).

Do Other States Use Stabilization?

Bayesian Stabilization Studies for NJ and PA

- The Regional Educational Laboratory (REL) Mid-Atlantic carried out a series of studies to use “Bayesian Stabilization” of school proficiency rates and growth rates (Rosendahl & Gill, 2024).
- This model leverages both data from within schools in prior years and from similar students at other schools.
- The model improved stability and reliability of small school/group estimates. The model worked well for indicators that were approximately normally distributed across schools (proficiency, growth) but not highly skewed indicators (5-year graduation rates).
- These stabilization methods **are not being used operationally** in either state.

EBLP Stabilization for California Growth Model

- California created a school/district growth model that was used for **reporting only** for the first time in early 2025.
- The growth metrics use “empirical best linear prediction” (EBLP) to stabilize estimates. This relies (almost entirely) on information from students in the same school in the prior year (McCaffrey & Castellano, 2021). The model produces a weighted average growth score across two years of data.
- CA uses a “hybrid” model – schools or districts with 500+ students report a simple average growth score; N<500 reports a stabilized score.
- CA uses a residual gain score growth model; minimum N-size of 11; no student-level results reported to anyone.

Value Added Models (VAM)

- Many value added models (VAM) have stabilization built into the model. Hence VAM school growth estimates may be stabilized. Because this is often built into the models, there is not necessarily a difference between stabilized and non-stabilized estimates in these models (although there can be).
- This is not directly relevant to CO context using SGP.
- Noting this to emphasize that while we have not found instances where states use stabilization of school average test scores or MGP for accountability, the framework of stabilization is built into some metrics.

CO EBLP Stabilization Study Terminology

- **Direct Estimate**: The observed mean scale score or mean SGP for in the current year. (The “simple average.”)

- **Stabilized Estimate**: The stabilized mean scale score or mean SGP in the current year that leverages data from current and prior year using EBLP.

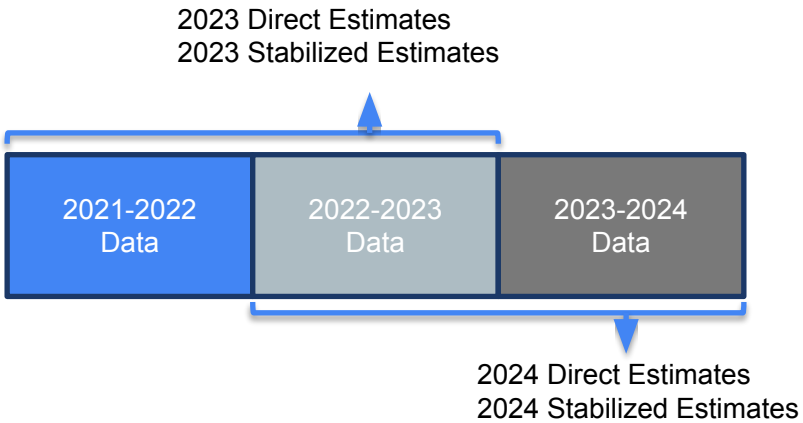
Research questions

- To what extent does EBLP improve stability of estimated mean SGP and mean scale scores?
- How much does use of EBLP stabilized estimates in place of direct estimates impact overall SPF ratings?
- How do these vary across grade levels (E,M,H), sample size, and subjects (math and ELA)?

Empirical Best Linear Predictor (EBLP)

- Developed by researchers at ETS in partnership with California. The stabilized metrics are computed as a weighted average of the unit's current year growth metric and the same unit's prior year growth metric. The weights are data-based, with relatively more weight on prior years for small units. In very large units, almost all weight is on current year.
- Key points:
  - Research in CA and another state show EBLP stabilized estimates are less variable year-to-year provide more accurate estimates of the true population quantity.
  - Can be **applied to means but not medians** (mean scale score or the mean SGP).
  - Theoretically superior to moving average (combine all scores across two years and average).

Study Overview



Calculate direct and stabilized estimates at the school-by-EMH level for all students and for each disaggregated subgroup with sufficient data for inclusion in SPF (ML, IEP, FRL, SOC). ***Focus on Mean SGP*** *(carry out similar analysis for mean scale scores).*

For each school with 2023 and 2024 estimates:

- **year-to-year correlation** of direct versus stabilized estimates.
- **year-to-year change** in the direct versus stabilized estimates.
- differences by **school size** and **level** (EMH).

For each school with 2024 data:

- **average accuracy ratio** between each school's direct and stabilized estimate.
- **average difference** between each school's direct and stabilized estimate.
- **agreement between SPF** results using direct versus stabilized estimates.
- differences by **school size** and **level** (EMH).

Accuracy Ratio

- The mean squared error (MSE) quantifies the uncertainty in an aggregate test score metric. In addition to calculating the year-to-year correlations and differences for the direct and stabilized metrics, we can evaluate the anticipated improvement in accuracy:

$$AccuracyRatio = \frac{MSE\ Direct}{MSE\ Stabilized}$$

- Values greater than 1.0 indicate the stabilized estimates are more accurate on average.

Detour: MeanSGP have higher year-to-year correlations and smaller average year-to-year changes than MedianSGP

					Year-to-Year Correlation		Average Year-to-Year Difference		2024 SD	
EMH	Group	Subject	Size	Schools	Median SGP	Mean SGP	Median SGP	Mean SGP	Median SGP	Mean SGP
EM	ALL	ELA	20-49	179	0.37	0.42	12.45	8.29	13.84	9.77
EM	ALL	ELA	50-149	889	0.47	0.50	8.82	6.00	11.16	7.89
EM	ALL	ELA	150+	466	0.59	0.61	6.70	4.66	9.49	6.77

- School **MeanSGP** are more highly correlated across years than school **MedianSGP**.
- School **MeanSGP** differ less year-to-year than school **MedianSGP**.
- School **MeanSGP** are less variable across schools within the same year than **MedianSGP**.
- Similar pattern for High Schools and across subjects.

Stabilization has greater impact on smaller groups (SGP)

EMH	Subject	Size	School/ Groups	Year-to-year Correlation			Average year-to-year difference				Avg. Accuracy Ratio
				Direct	Stabilized	Diff.	Direct	Stabilized	Diff.	MAD	
EM	ELA	20-49	1523	0.39	0.64	*0.25*	7.45	5.15	*-2.30*	2.15	1.43
EM	ELA	50-149	2319	0.49	0.62	*0.14*	5.92	4.78	*-1.14*	1.00	1.21
EM	ELA	150+	844	0.60	0.66	*0.06*	4.75	4.28	*-0.47*	0.39	1.09
H	ELA	20-49	312	0.36	0.68	*0.32*	6.86	4.32	*-2.54*	2.34	1.58
H	ELA	50-149	396	0.53	0.75	*0.22*	4.82	3.19	*-1.63*	1.20	1.37
H	ELA	150+	396	0.69	0.80	*0.11*	3.07	2.42	*-0.65*	0.46	1.17

MAD=mean absolute difference between direct and stabilized.

- Table includes school overall and disaggregated groups results.
- Results similar across subjects and for each disaggregated group.

Stabilization has greater impact on smaller groups (scale scores)

EMH	Subject	Size	School/ Groups	Year-to-year Correlation			Average year-to-year difference				Avg. Accuracy Ratio
				Direct	Stabilized	Diff.	Direct	Stabilized	Diff.	MAD	
EM	ELA	16-29	1017	0.91	0.95	*0.05*	6.92	4.71	*-2.21*	2.80	1.44
EM	ELA	30-149	3526	0.93	0.96	*0.03*	5.03	3.84	*-1.20*	1.47	1.25
EM	ELA	150+	1475	0.97	0.98	*0.01*	3.42	3.10	*-0.32*	0.57	1.12
H	ELA	16-29	206	0.87	0.96	*0.09*	20.97	12.51	*-8.46*	10.58	1.84
H	ELA	30-149	650	0.95	0.98	*0.03*	14.98	10.86	*-4.13*	5.98	1.59
H	ELA	150+	460	0.98	0.99	*0.01*	8.82	6.67	*-2.15*	2.86	1.39

MAD=mean absolute difference between direct and stabilized.

- Table includes school overall and disaggregated groups results.
- Results similar across subjects and for each disaggregated group.

Summary

- EBLP stabilized estimates have **higher year-to-year correlations** and **smaller year-to-year differences** relative to direct estimates.
- These differences were **more pronounced for smaller groups** and for mean SGP versus mean scale scores (both as expected).
- Similar pattern of results observed in prior studies of EBLP.
- Report with more detailed comparisons broken out by groups, subjects, etc. in preparation. Pattern of results remains consistent; stabilization may have greater impact on high school.

Overall School Ratings Impact: Direct vs Stabilized Estimates

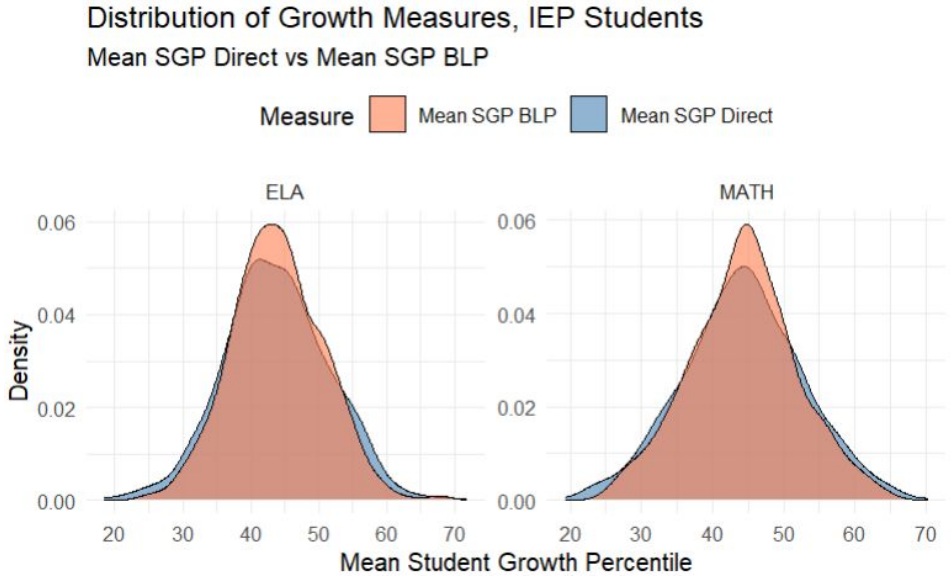
	Stabilized Estimate			
Direct Estimate	**Performance**	**Improvement**	**Priority Improvement**	**Turnaround**
Performance	**1246**	**36**	0	0
Improvement	**35**	**338**	**12**	**1**
Priority Improvement	**1**	**28**	**140**	**8**
Turnaround	0	0	**9**	**88**

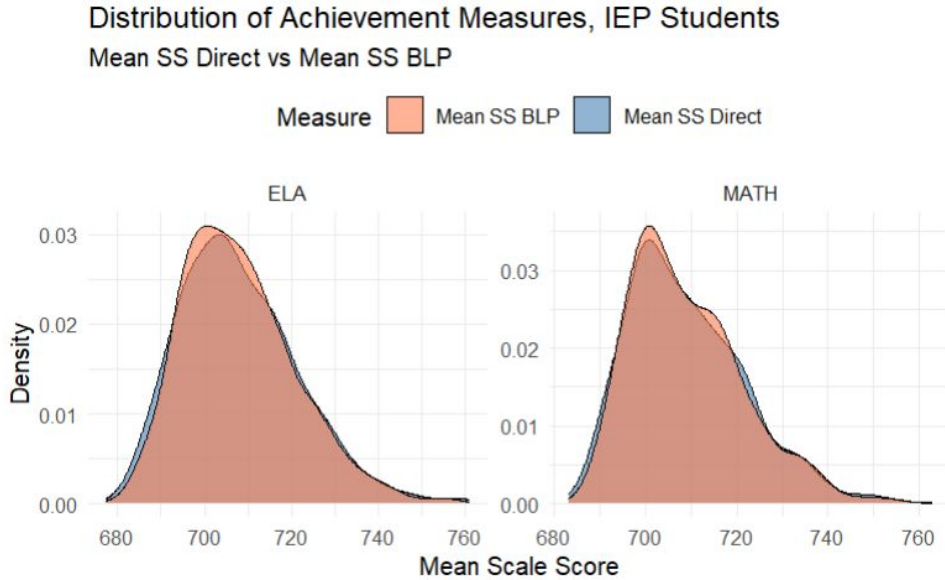
Summary of Ratings Impact

Change	Count	Percent
Up 2	1	0.1%
Up 1	72	3.7%
No Change	1812	93.3%
Down 1	56	2.9%
Down 2	1	0.1%

- 2024 SPF data only, excl. AECs
- Minimum N-sizes remain 16 (achievement) and 20 (growth).
- After stabilizing both achievement and growth measures, 93% of school ratings do not change.
- ~3% of school ratings move downwards
- ~4% of school ratings move upwards

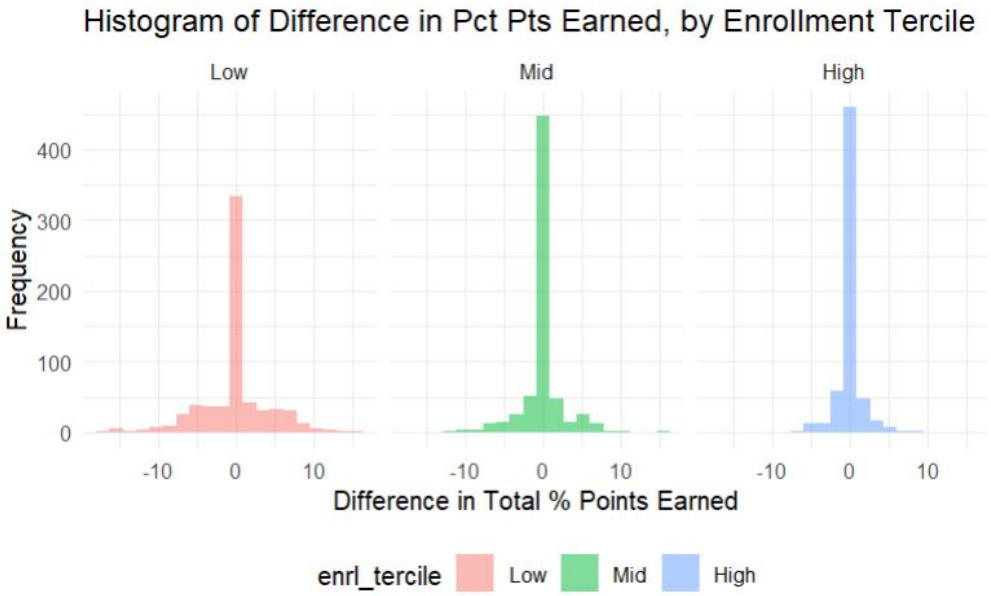
Direct vs Stabilized Growth and Achievement Distributions: Subindicator Level, Students with IEPs





- Switching from the direct mean to the stabilized BLP mean has a relatively small impact on growth and achievement measure distributions, pulling data in tails slightly towards center
- Again, smaller groups/schools will experience more shifts in measures

Change to Total % of Framework Points Earned, by Enrollment Tercile



Lowest Tercile	< 247 students
Middle Tercile	247 to 430 students
High Tercile	> 430 students

- Looking by enrollment size, we see that smaller schools experience more frequent and greater magnitude shifts in total % of points earned.
- For large schools, changes are mostly less than 5% pts.
- For small schools, changes can be more than 15% pts in some cases.

Overall School Ratings Impact, by Enrollment Tercile

High Enrollment Tercile

Change	Count	Percent
Up 1	7	1.1%
No Change	603	97.4%
Down 1	9	1.5%

Mid Enrollment Tercile

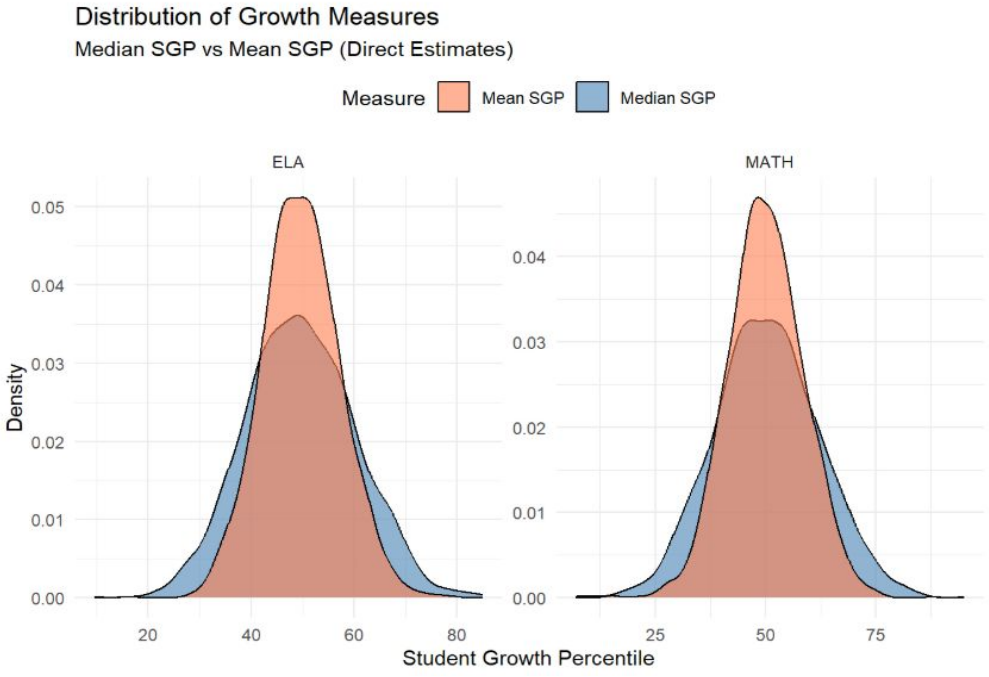
Change	Count	Percent
Up 1	20	3%
No Change	619	93.5%
Down 1	23	3.5%

Low Enrollment Tercile

Change	Count	Percent
Up 2	1	0.2%
Up 1	45	6.8%
No Change	590	89.3%
Down 1	24	3.6%
Down 2	1	0.2%

- As expected, schools in the lowest enrollment tercile experience more frequent ratings shifts as a result of stabilization, with more schools moving up than down.
  - 97% of high enrollment schools retain the same rating
  - 89% of low enrollment schools retain the same rating

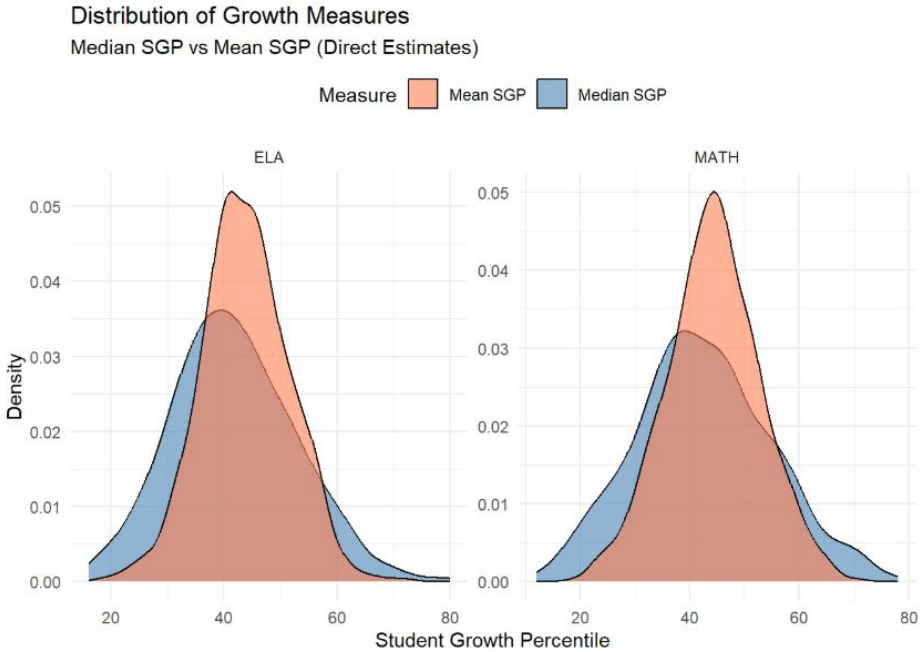
Median vs Mean SGP Distributions: Subindicator Level, All Students



- Comparing ratings when shifting from median SGPs to mean SGPs involves important operational decisions. Further study is needed to determine best approach.
- For initial analysis, distribution of mean SGP was renormed to find 15th/50th/85th percentile cuts for each subject and grade level.
- Example below shows how cutscore move towards middle of distribution.

subject	emh	target_pctl	cutscore	n_sch
ELA	E	15	41.97	1010
ELA	E	50	49.91	1010
ELA	E	85	58.55	1010

Median vs Mean SGP Distributions: Subindicator Level, Students with IEPs



- Median SGPs for smaller disaggregated groups tend to be lower than for All Students.
- Switching to mean SGPs pulls the distribution to the right and narrows it.
- The results is that disaggregated group ratings are often lower once the distribution of means is renormed.

Overall Ratings Impact: Median SGP vs Mean SGP

Rating w/ Median SGP	Rating w/ Direct Mean SGP			
	Performance	Improvement	Priority Improvement	Turnaround
Performance	1264	85	0	0
Improvement	18	292	72	1
Priority Improvement	0	9	102	41
Turnaround	0	0	3	55

Summary of Ratings Impact

Change	Count	Percent
Up 1	30	1.5%
No Change	1713	88.2%
Down 1	198	10.2%
Down 2	1	0.1%

- Under current specifications, change from median to mean SGP results in substantial shift to overall ratings.
- 88% of schools stay the same
- 10% of schools move downwards, while only 1.5% move upwards.
- This shift is an artifact of our initial analytic strategy and would likely look different were we to thoroughly investigate using mean SGPs.

Takeaways

- Overall ratings impacts resulting from EBLP stabilization are modest
  - 3-4% of schools change ratings
  - Changes are balanced (up/down)
  - Smaller schools are impacted more than larger schools
- The change from median SGP to mean SGP would likely result in bigger shifts in ratings.
  - Methodology for such a change would need further study

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

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

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 (2022). schoolgrowth: Functions for Improving Accuracy of School-Level Growth Measures Using Empirical Best Linear Prediction. R package version 0.6-3, <https://github.com/jrlockwood/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

Rosendahl, M., & Gill, B. (2024, November 7). *A solution for promoting accuracy and equity in measures of school performance*. https://ies.ed.gov/learn/blog/solution-promoting-accuracy-and-equity-measures-school-performance

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

Overview

Background	• Legislative Background • Importance of Min N Size • National/Colorado context
Review Reliability Study	• Research Questions & Findings • See resources from May 2025 TAP meeting for more info
Introduce Stabilization Study	• Stabilization Goals and Context • EBLP Methodology • Research Questions & Findings
Discussion	• What are you taking away from these two studies? • Questions/concerns for follow up?

Public Comments & Meeting Close

Aislinn Wales & Scott Weldon

Technical Advisory Panel

- **Meeting Summary**
  - Suggested future analysis
  - TAP recommendations from this meeting
- **Public Comment**
- **Close Meeting**
  - Next Scheduled Meeting: October 23, 2025
  - Poll for December Meeting