COLORADO
Department of Education

Technical Advisory Panel Meeting

October 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**

Agenda for Today

- **Welcome and Introductions** | Information Item
- **2026 Performance Framework Targets** | Information Item
- **Small N Stabilization Study Background** | Feedback Item
- **Growth from CMAS to PSAT** | Feedback Item
- **Wrap-Up**

Update on 2026 Performance Framework Targets

Lisa Medler

Information 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

Sub-Indicator Targets | Options for State Board

#	Option	Description
1	**Hold:** Hold existing targets	Keep 2026 sub-indicator targets consistent with **2024 baseline**.
2	**Wait to Review/Revise:** Hold existing targets but plan to **update PSAT/SAT based on 2026 results** (if needed)	Keep 2026 sub-indicator targets consistent with **2024 baseline.** Review PSAT/SAT and **potentially revise targets** once **2026 data** become available. If appropriate, reconcile 1- and 3-year targets.*
3	**Increase and Review/Revise:** Hold existing targets but **update PSAT/SAT to 2025 baseline** and **update based on 2026** results (if needed)	Keep 2026 sub-indicator targets consistent with 2024 baseline, except for **PSAT/SAT** which will be **updated to match 2025**. Review and **potentially lower targets** once **2026 data** become available. If appropriate, reconcile 1- and 3-year targets.*
4	Other revision	CDE would need more guidance from the state board for a new methodology.

* Targets will be revised (as necessary) using 15-50-85th percentile methodology

Update on State Board Decision Making

- Presented to information item to State Board on October 9. Scheduled to vote on November 12-13.

- They asked for feedback from additional stakeholder groups

- CDE staff are working with
  - ACEE
  - AWG
  - Rural Council
  - Accountability Contacts
  - 1x1s with State Board members

Small N Stabilization Study

Daniel Mangan

Feedback Item

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

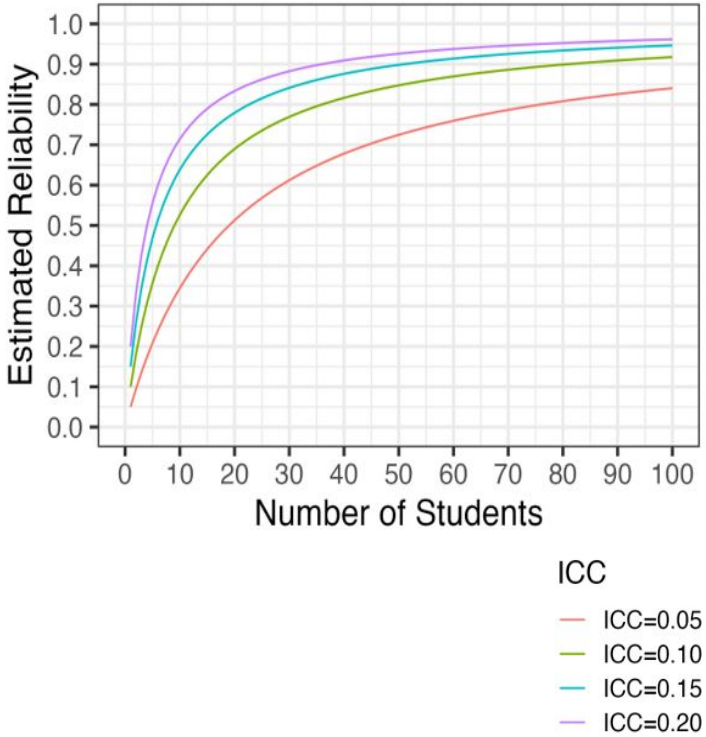
CRS 22-11-212(1)(a) and (2)(a)

(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**.

Review: Reliability Study

- 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.
- 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 20 pct point increase in share of schools w/ reportable data schools. (In the case of Elementary IEP students, up to 45 pct pt increase)
- 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.



National Context

- Three states average across three years for all accountability purposes (AK, WA, VT).
- Of the states that use minimum N = 10 for accountability, three states aggregate across multiple years to reach the minimum N size (ND, SD, OK).
- Only CO and NM use different minimum N sizes for growth and achievement.

(Alliance for Excellent Education, 2018)

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.

Population vs. Sampling Models

- Cronbach (1997) specified two perspectives from which to consider a population of interest (POI).
- Framework calculations are built fundamentally on a finite population model, but some proposed methodologies require a sampling (infinite population) approach.
- Infinite population model is supported by AERA, APA, and NCME standards.

Finite Population (Population)

- POI = specific students in specific year
- Only measurement error contributes to uncertainty (no sampling error)

Infinite Population (Sample)

- POI = random sample of students from the population associated with school/group
- Sampling error and measurement error both contribute to uncertainty

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 averages).

Stabilization Study Takeaways

- EBLP stabilized estimates have higher year-to-year correlations and smaller year-to-year differences relative to direct estimates.
  - Differences more pronounced for smaller groups and for mean SGP vs mean scale score (both as expected)
- Overall ratings impacts resulting from EBLP stabilization are modest
  - 2-5% 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

Reccomendation 1: Addressing Measure Volatility

Option A: No Change

- Do not adopt EBLP stabilization
- Retain use of median SGP

Option B: Adopt mean SGP, no Stabilization

- Use mean SGP instead of median
- Phase-in w/ informational year

Option C: Adopt mean SPG + EBLP Stabilization

- Use EBLP for stabilization of mean SGP and mean SS
- Phase-in w/ informational year, publish side-by-side results

Reccomendation 2: Lowering Minumum N Thresholds

Option A: No Change

- Retain minimum N thresholds of 16 for achievement and 20 for growth

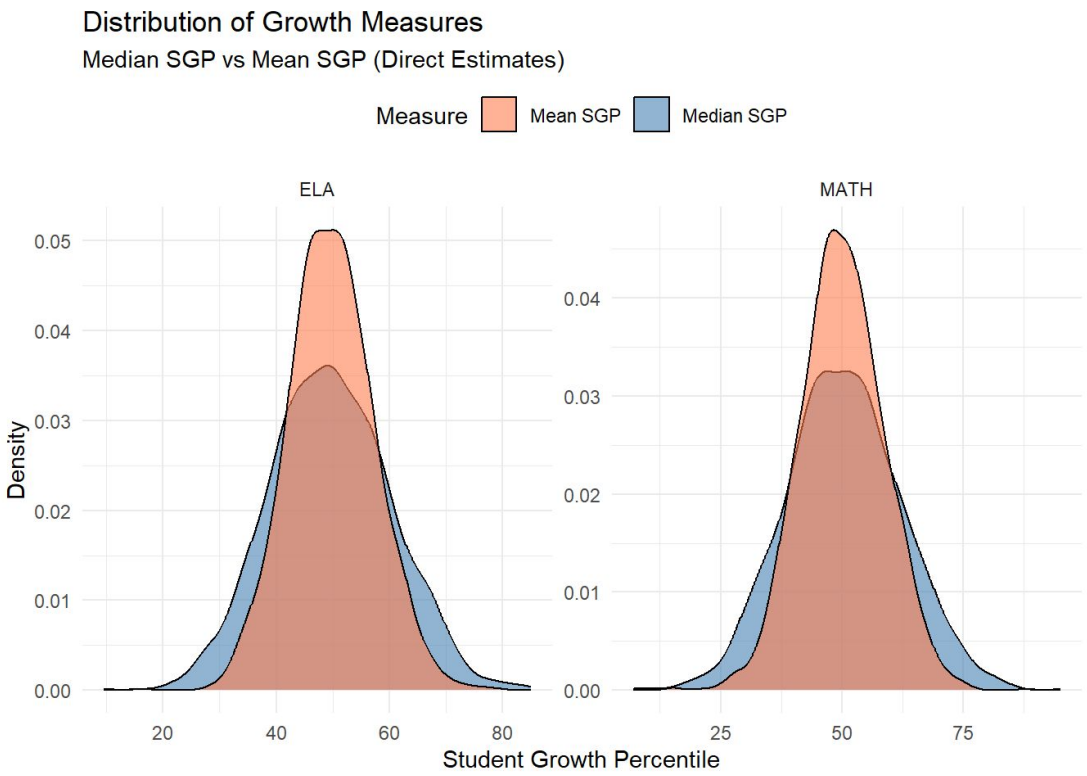
Option B: Lower Min N Size, Same N for ACH/GRO

- Allow n = 10 for both achievement and growth to expand visibility
- Quantify gains in visibility

Option C: Lower Min N Size, Different N for ACH/GRO

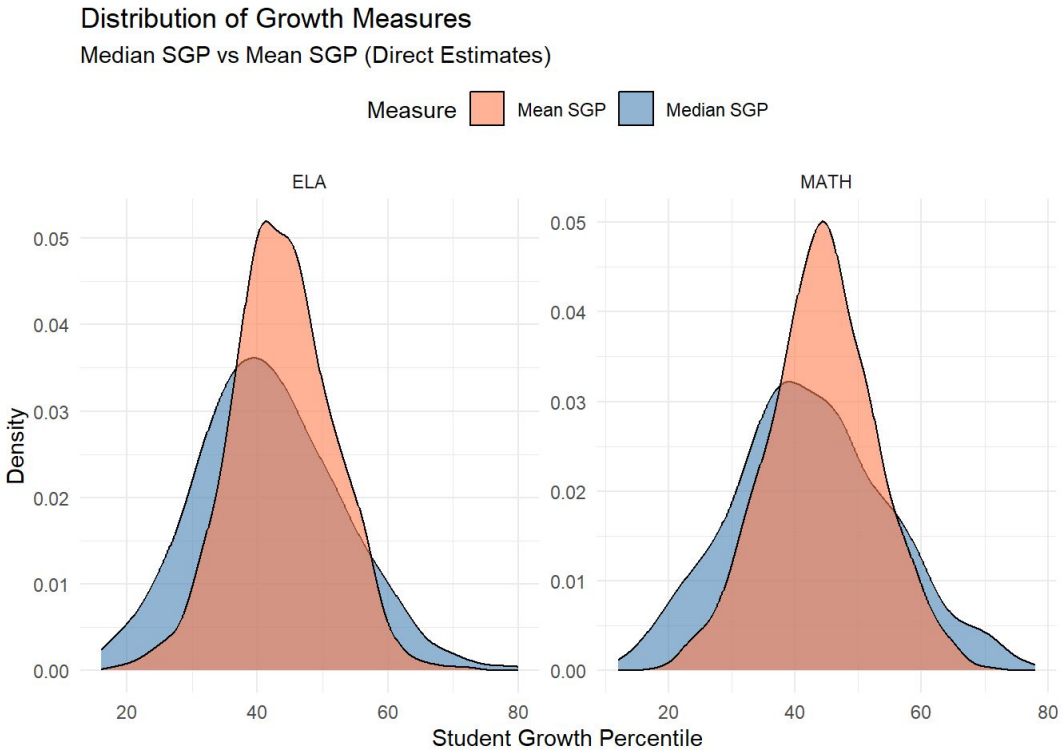
- Lower n count thresholds but continue to differentiate for achievement and growth (e.g…..12/15)
- Consider differentiation for reporting/accountability

Mean SGP vs Median SGP
ALL Students



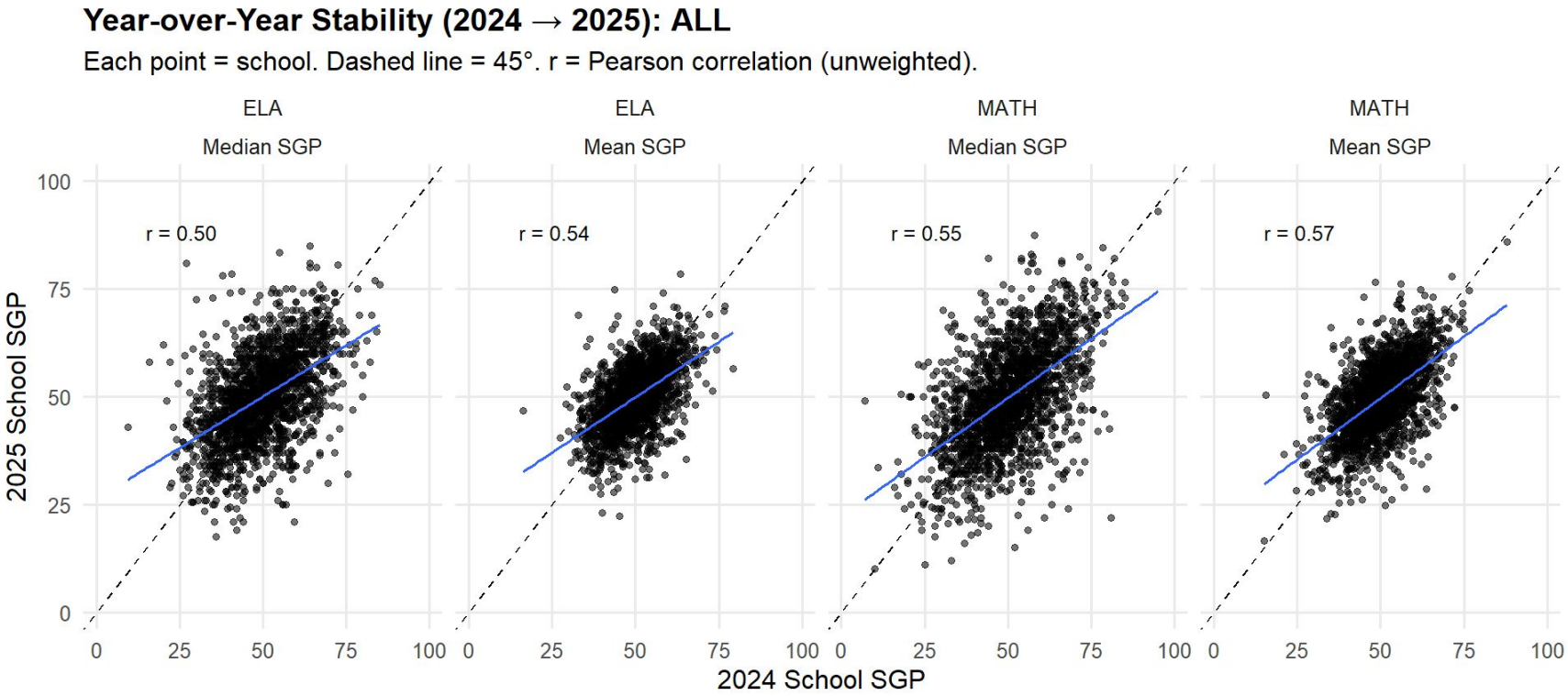
- Mean SGPs…
  - …are less variable across schools because *all* student scores are being used and random variation cancels out.
  - …provide a more stable signal across years (lower standard error) but can be swayed by extreme values
- Median SGPs…
  - …are more robust to outliers but also less precise, particularly for small samples.
  - …capture a school's "typical" student but can yield volatile results for small systems.

Mean SGP vs Median SGP
IEP Students

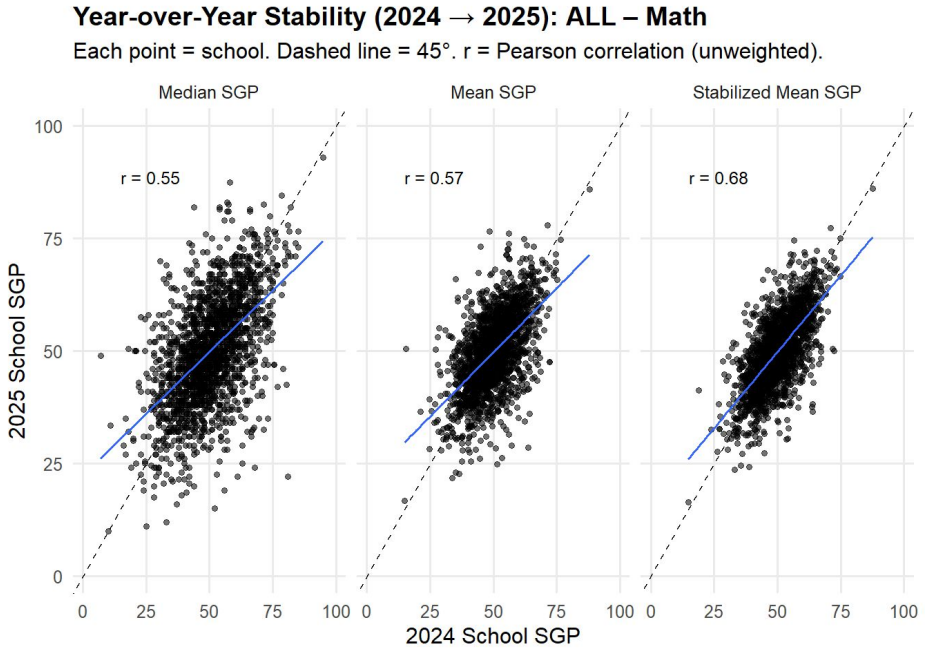


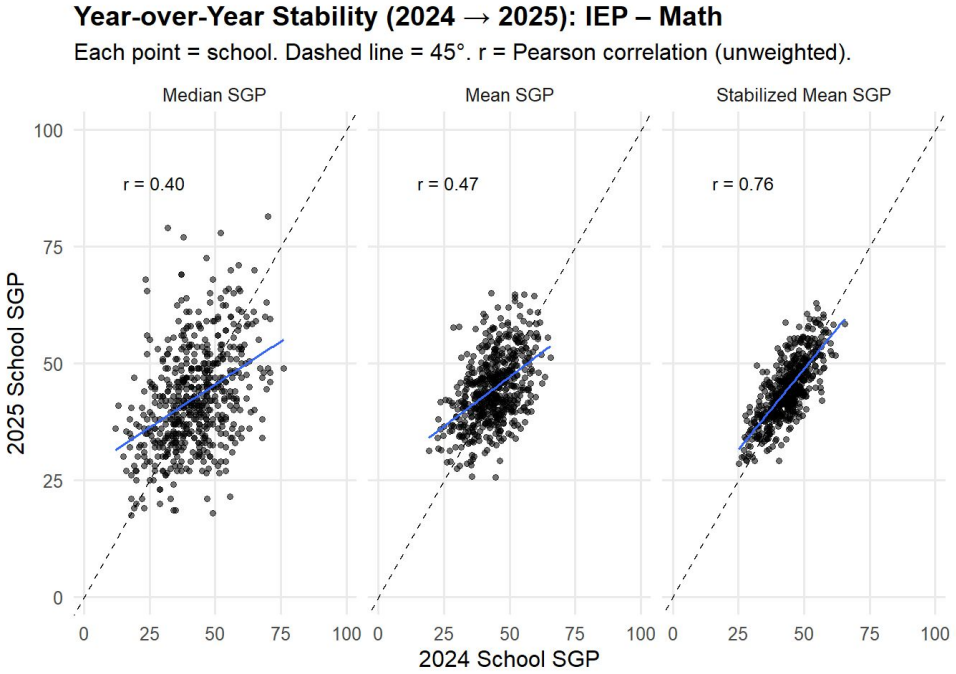
- For disaggregated student groups, particularly IEP students, median SGP distributions tend to be **shifted left** and have a slight **right-skew**.
- The median mitigates the weight of clustered low SGPs within each school, but across schools it behaves noisily because of small-N discreteness and floor effects.
- Because low-growth students are fully represented in the mean, overall subgroups averages can shift downward, causing more schools and groups like IEP students to fall below the growth rating cut points.

Mean SGPs are More Stable Year-toYear than Medians



EBLP Stabilized Mean SGPs Substantially Increase Year-to-Year Correlations, Esp. for Small Groups





Important Clarification: What Volatility Do We Want to Address?

- It is not clear from the 1241 TF or statute which volatility is of greatest concern:
  - Volatility in aggregate growth measures can be addressed by using mean SGP instead of median.
  - Volatility in both growth and achievement measures can be further addressed with empirical stabilization.
  - However, due to the normative nature of ratings, this won't necessarily lead to fewer changes in ratings.

Mean SGP Target Setting

- Growth rating cut points were originally based on *percentiles* of the statewide *median SGP* distribution.
  - 15/50/85 percentiles roughly corresponded to Median SGPs of 35/50/65
- But because the **mean SGP distribution is narrower**, the same percentiles correspond to **different score values**.

	15th Pctl	50th Pctl	85th Pctl
Median SGP	~35	~50	~65
Mean SGP	~40	~50	~60

- With a compressed distribution, smaller shifts in mean SGP can result in ratings changes because the cut points are closer together.
- Although mean SGPs are more precise, the narrower distribution could result in ***increased*** rating churn.
- Stabilization would likely mitigate some of this increased churn.

2024 – 2025 Ratings Transition Matrix
Original Overall Ratings, All Schools

- Using original ratings (median SGP, no stabilization):
  - 74.5% of schools retain the same rating
  - 14.2% of schools increase ratings
  - 11.4% of schools decrease ratings

Rating Comparision, 2024 to 2025

Original Ratings (using Median SGP, no stabilization)

	2025 Rating			
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

Change	Count	Percent
Up 3	5	0.3%
Up 2	47	2.5%
Up 1	217	11.4%
No Change	1418	74.5%
Down 1	184	9.7%
Down 2	31	1.6%
Down 3	2	0.1%

2024 – 2025 Ratings Transition Matrix
Direct Overall Ratings, All Schools

- Using direct ratings (mean SGP and SS, no stabilization):
  - 71% of schools retain the same rating
  - 15.8% of schools increase ratings
  - 13.2% of schools decrease ratings

Rating Comparision, 2024 to 2025

Direct Ratings (using mean SGP, no stabilization)

	2025 Rating			
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

Change	Count	Percent
Up 3	4	0.2%
Up 2	38	2%
Up 1	258	13.6%
No Change	1352	71%
Down 1	214	11.2%
Down 2	31	1.6%
Down 3	7	0.4%

2024 – 2025 Ratings Transition Matrix
Stabilized Overall Ratings, All Schools

- Using stabilized ratings (mean SGP and SS, EBLP stabilization):
  - 74.4% of schools retain the same rating
  - 13.2% of schools increase ratings
  - 11.5% of schools decrease ratings

Rating Comparision, 2024 to 2025

Stabilized Ratings

	2025 Rating			
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

Change	Count	Percent
Up 3	1	0.1%
Up 2	25	1.3%
Up 1	244	12.8%
No Change	1417	74.4%
Down 1	190	10%
Down 2	26	1.4%
Down 3	1	0.1%

Within-Year Ratings Transition Summary
2024 Overall Ratings, All Schools

	2024 Original - 2024 Direct Ratings				2024 Direct - 2024 Stabilized Ratings			
	%UP	%SAME	%DOWN	DIFF	%UP	%SAME	%DOWN	DIFF
All Schools	1.9	92	6.1	-4.2	3.4	93.8	2.9	0.5
Large	0.8	93.1	6.1	-5.3	1.6	97.2	1.3	0.3
Medium	2.1	91.7	6.2	-4.1	3.9	92.6	3.5	0.4
Small	3.9	90.1	6	-2.1	5.4	90.4	4.2	1.2
Elementary	2.6	90.5	6.9	-4.3	4.2	92.3	3.6	0.6
Middle	2	92.7	5.4	-3.4	3.4	94.5	2.1	1.3
High	0	94.9	5.1	-5.1	1.3	96.8	1.9	-0.6

- Looking within a single year, moving from median SGP to mean SGP (with current method) changes overall rating for 5-10% of schools, with a small aggregate downward shift.
- Comparing single year direct mean SGP to stabilized mean SGP ratings, we see an overall shift upwards, but with less magnitude than the median-mean shift.
- The overall downwards shift when moving from median to mean SGP is likely due to the right-skew of median SGP distributions for many subgroups.

Enrollment Groups: Small < 150; 150 <= Medium < 400; 400 <= Large

Year-over-Year Ratings Transition Summary 2024-2025 Overall Ratings, All Schools

	2024 - 2025 Original Ratings				2024 - 2025 Direct Ratings				2024 - 2025 Stabilized Ratings			
	%UP	%SAME	%DOWN	DIFF	%UP	%SAME	%DOWN	DIFF	%UP	%SAME	%DOWN	DIFF
All Schools	14.2	74.5	11.4	2.8	15.8	71	13.2	2.6	14.3	74.4	11.5	2.8
Large	10.7	80.5	8.8	1.9	13	77.5	9.5	3.5	12.1	79.1	8.8	3.3
Medium	15.9	70.3	13.8	2.1	16.8	67.1	15.9	0.9	14.8	71.8	13.4	1.4
Small	16.5	72.6	10.8	5.7	18.7	67.2	14.1	4.6	16.9	71.4	11.7	5.2
Elementary	15.7	70.8	13.6	2.1	17.3	66.6	16.3	1	15.1	71.3	13.7	1.4
Middle	15.9	72.4	11.7	4.2	16.5	70.6	13	3.5	14.1	74.6	11.3	2.8
High	7.2	87.8	5	2.2	10.5	83.9	5.5	5	11.9	82.8	5.3	6.6

- Compared to within-year rating shifts related to either mean-median change or stabilization, year-over-year ratings show substantially churn
  - 2024-25 Original Ratings: 12-30% of school ratings change
  - 2024-25 Direct Ratings: 16-34% of school ratings change
  - 2024-25 Stabilized Ratings: 17-28% of school ratings change

Enrollment Groups: Small < 150; 150 <= Medium < 400; 400 <= Large

Feedback on Options for Stabilization and Reducing Min N Size

- The linked feedback document has separate tabs for each of the two recommendations.
- We will be asking the TAP to decide on recommendations in the November meeting.
- For today, please use this document to process your thinking around the available options for both recommendations:
    - Add Pros and Cons to the list for each option
    - Note additional considerations that are not “pros” or “cons”
    - Pose questions and/or request follow-up analysis

Growth from CMAS to PSAT

Marie Huchton

Feedback Item

Summary of 2019 Investigation

- Under the previous paper-based PSAT assessment, CDE found that the grade 8 CMAS ELA construct was not sufficiently aligned to the grade 9 PSAT Reading & Writing (RW) construct to produce usable student growth percentile calculations.
- Summary of Findings from 2019 investigation:
  - Lower between-year correlations (related to gender differences in the student performance trends from CMAS ELA to PSAT EBRW) as compared to the other grade and content areas.
  - Attributed to the PSAT/SAT not including a substantial writing component (selected response items only) while CMAS ELA includes explicit text-based writing tasks (constructed response items in addition to selected response items).

Summary of 2019 Investigation (cont.)

- CMAS ELA was designed to measure Colorado's academic content standards, including a student's ability to produce original written work.
- The PSAT EB/RW construct does not align as closely with Colorado's academic content standards, including not requiring original student writing tasks, which was an acknowledged part of the trade-off in moving to PSAT/SAT.
- Achievement and growth results for girls taking Colorado's ELA assessments have historically been higher than for boys, particularly for CSAP/TCAP Writing and CMAS ELA, likely due to the inclusion of open-ended constructed response items.

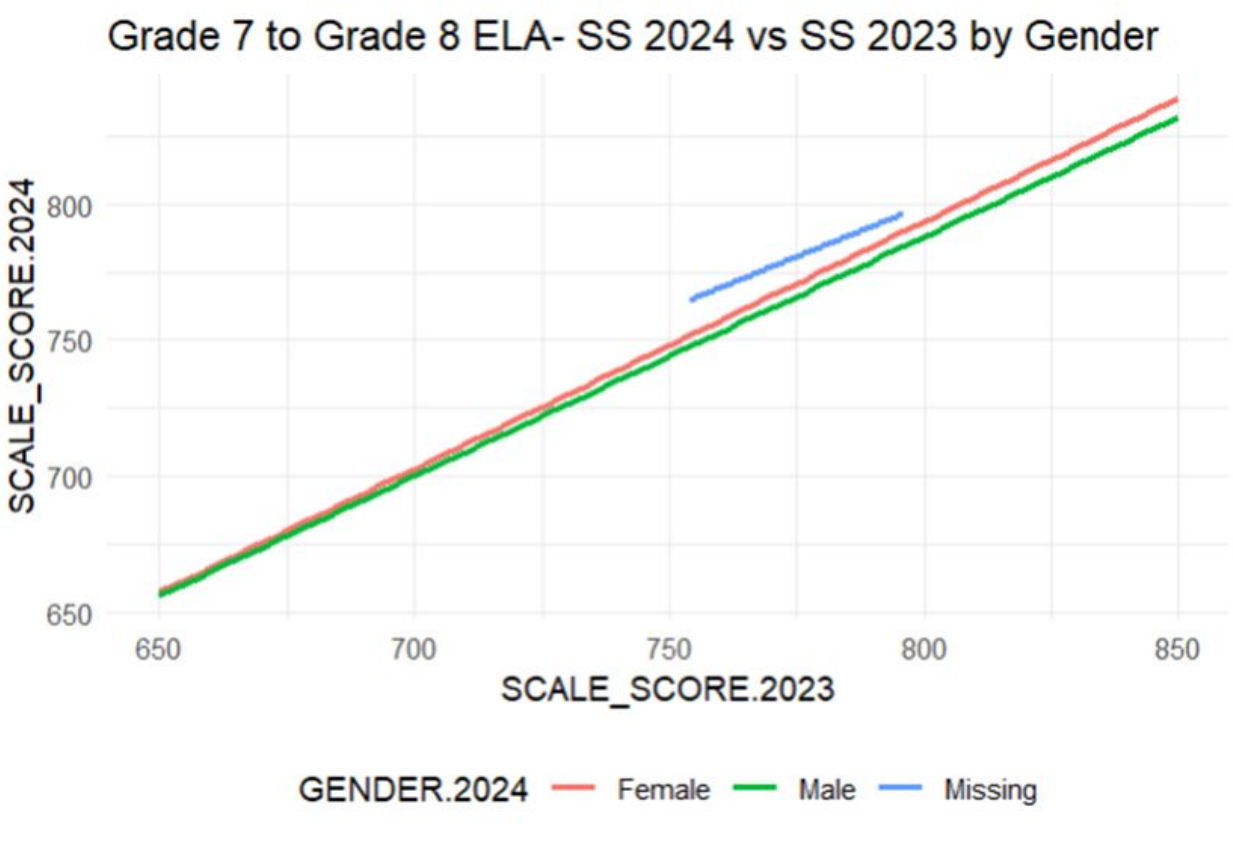
Summary of 2019 Investigation (cont.)

- As students transition from CMAS ELA to PSAT EBRW, the relative student rankings shift in different ways for boys and girls due to the different assessment constructs.
- An unfortunate side effect of Colorado's normative growth model calculation is that any systematic differences in score rankings between years are magnified in the student growth percentile results.
- TAP reviewed the 2018 and 2019 scale score and growth results from CMAS g8 ELA to PSAT g9 EBRW for girls and boys against the other grades/content areas and supported the decision to not run growth for grade 9 EBRW.

Updated Investigation for 2025

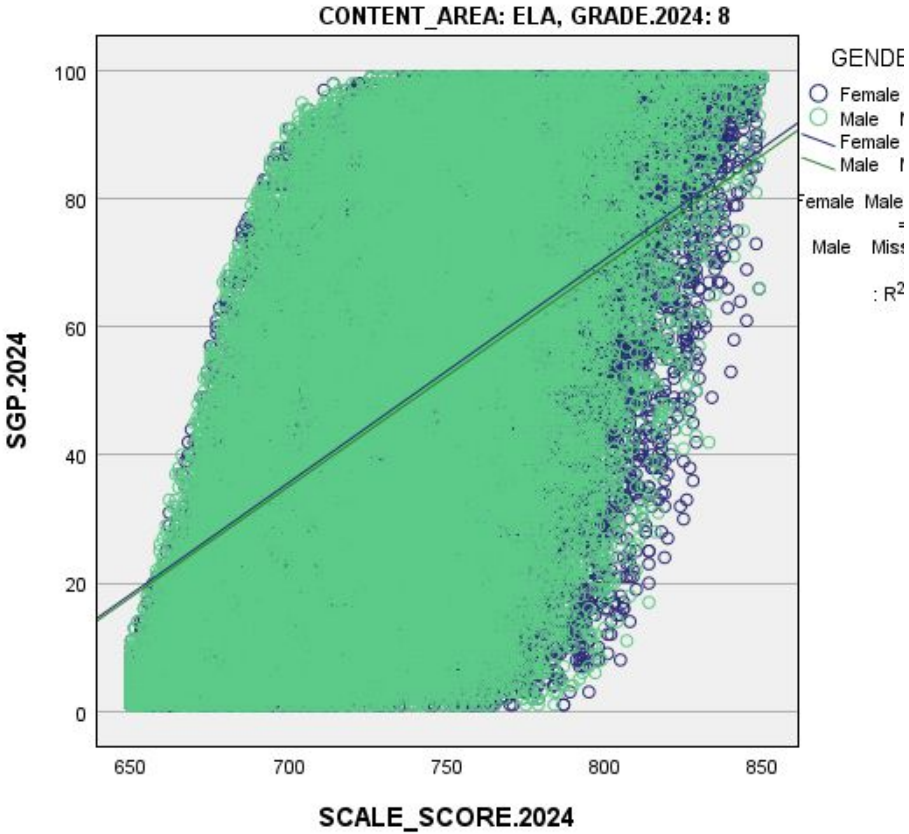
- Now that we have two years of the new digital PSAT assessment, we can reassess linking ELA and RW for growth
- Considerations of note for the digital assessment:
  - Still only contains selected-response items
  - Passages were reduced in length and include only one item per passage rather than previous multiple items per passage

Scale Score Trends from 2023 to 2024: CMAS g7 to CMAS g8 ELA by Gender



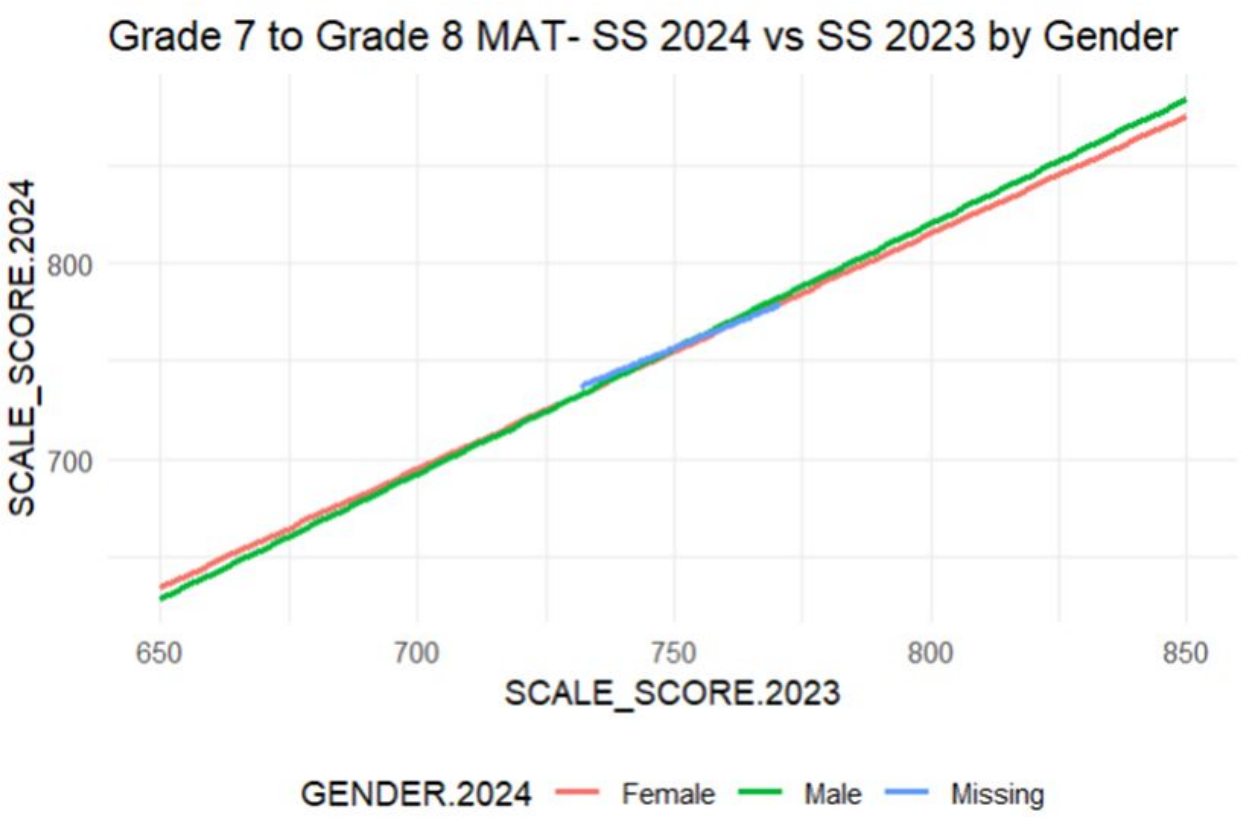
Female students show slightly larger scale score increases between years than male students, especially at higher levels of historic performance.

SGP to Scale Score Scatter plot for 2024: CMAS g8 ELA by Gender



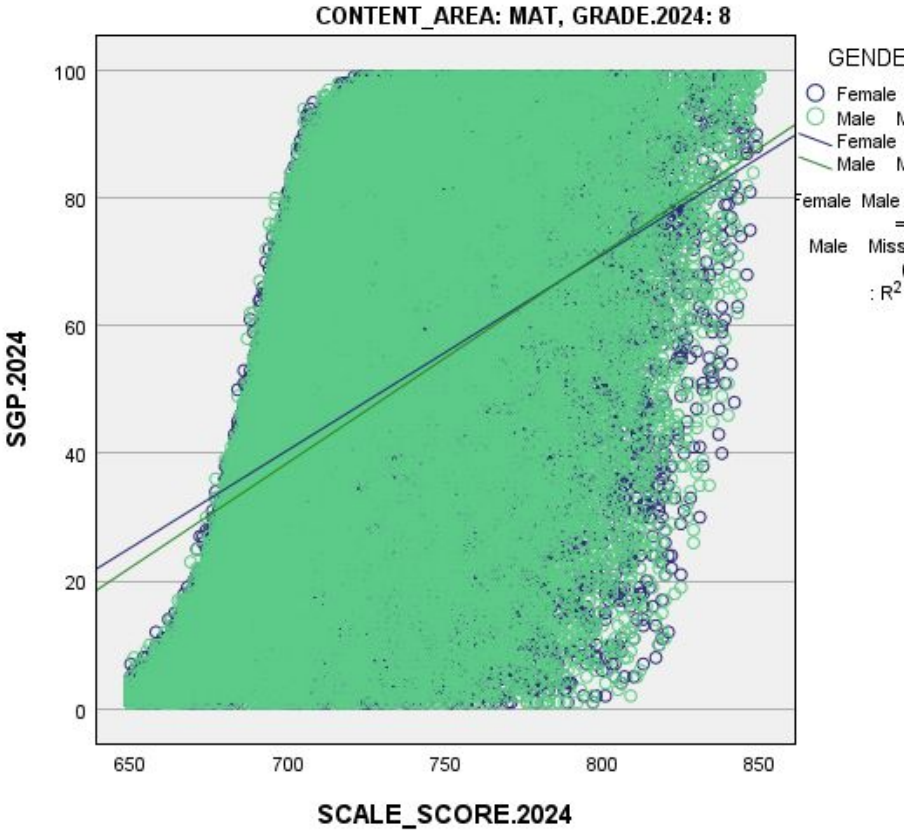
Aligning with the scale score trends over time, female students show slightly higher growth percentiles than male students.

Scale Score Trends from 2023 to 2024: CMAS g7 to CMAS g8 Math by Gender



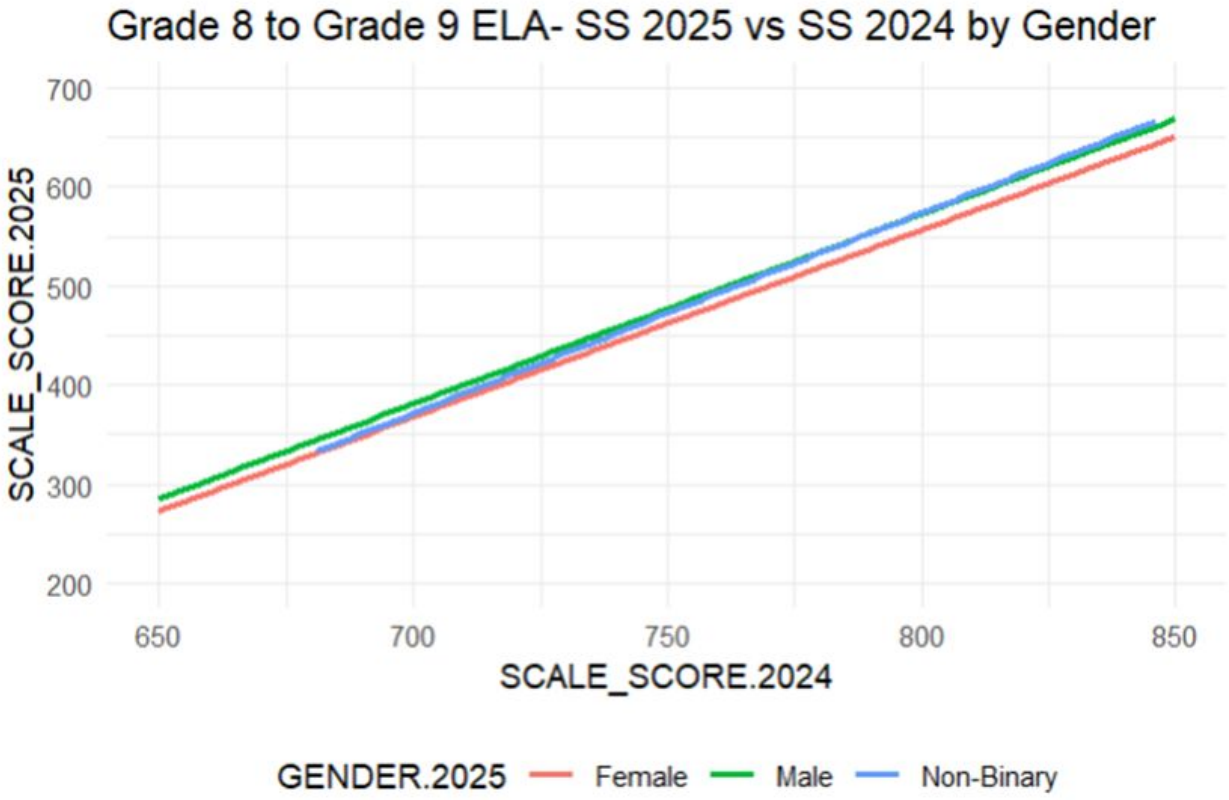
Historically lower performing male students show slightly smaller scale score increases than female peers, but historically higher performing males show slightly larger score increases than female peers.

SGP to Scale Score Scatter plot for 2024: CMAS g8 Math by Gender



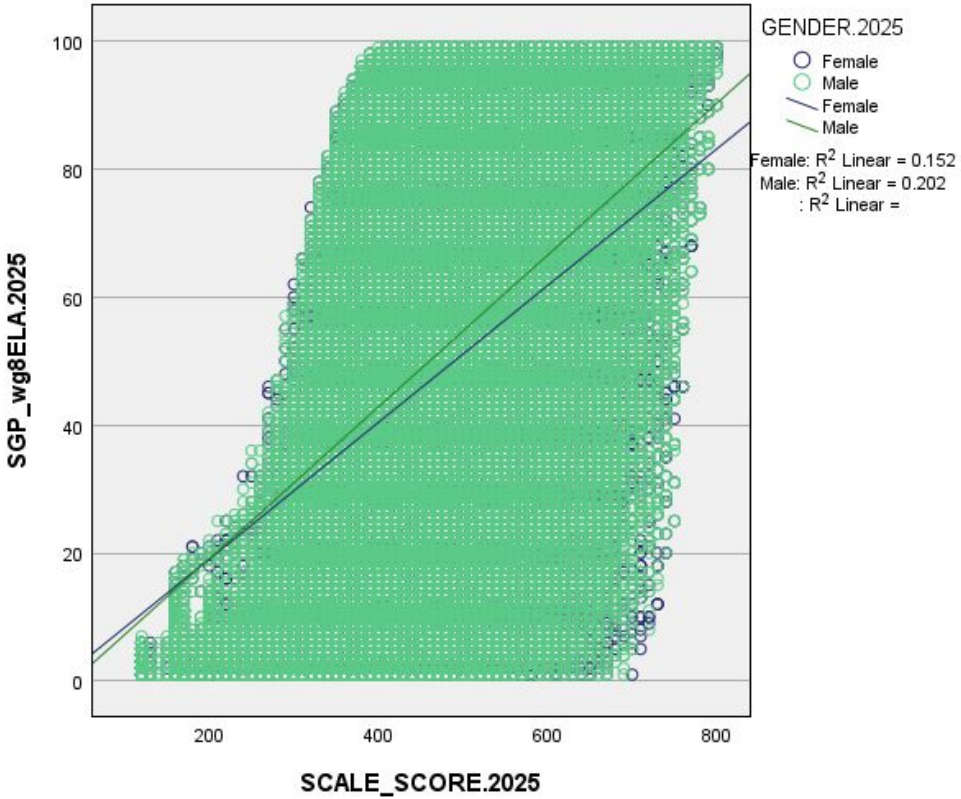
Slightly exaggerating the trends from the scale scores, historically lower performing male students show lower growth than female peers, but historically higher performing males show slightly larger score increases than female peers.

Scale Score Trends from 2024 to 2025: CMAS g8 ELA to PSAT g9 RW by Gender



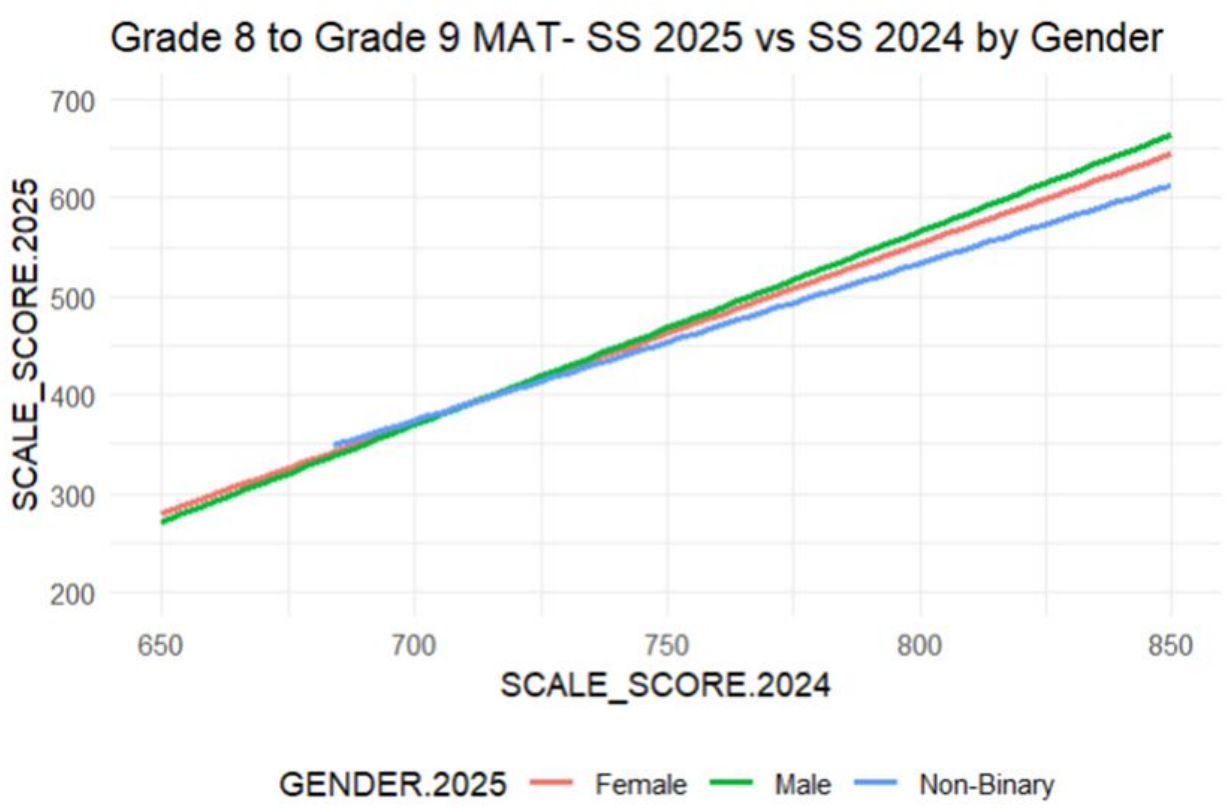
Males show slightly larger scale score increases than female peers. This is true across all levels of historical performance, but more substantial for higher performers.

SGP to Scale Score Scatter plot for 2025: PSAT g9 ELA/RW by Gender



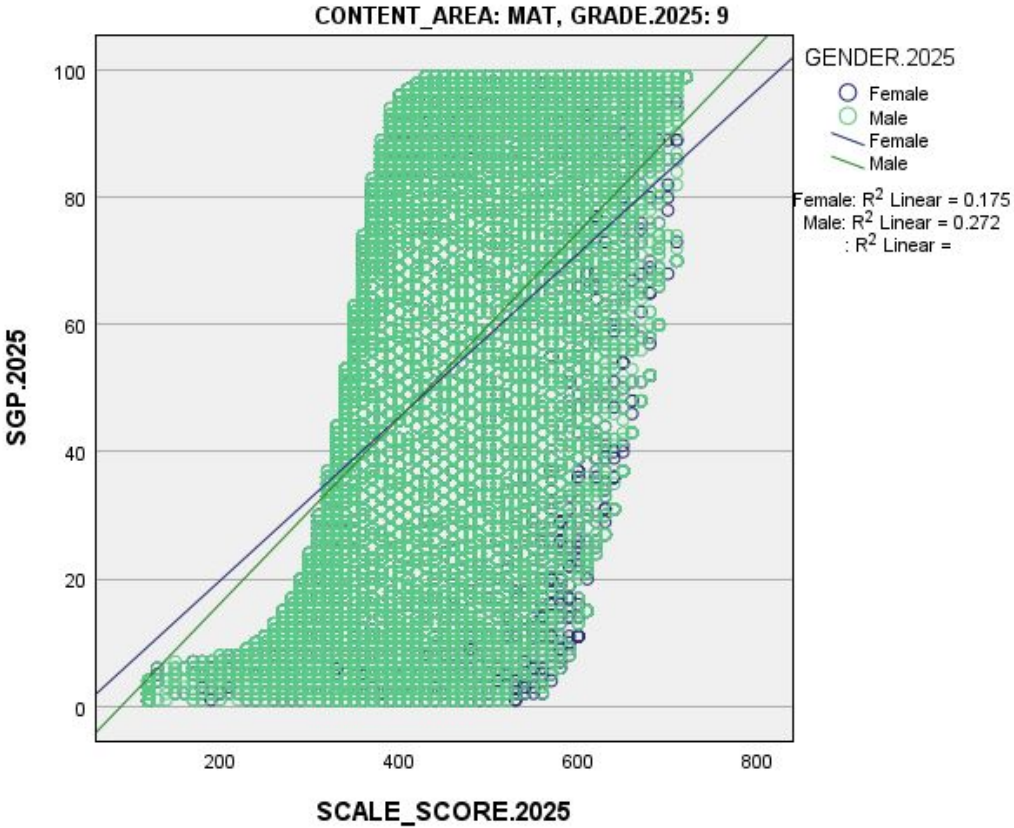
Partially aligned with the trends seen for scale scores, historically higher performing males show higher SGPs than female academic peers.

Scale Score Trends from 2024 to 2025: CMAS g8 to PSAT g9 Math by Gender



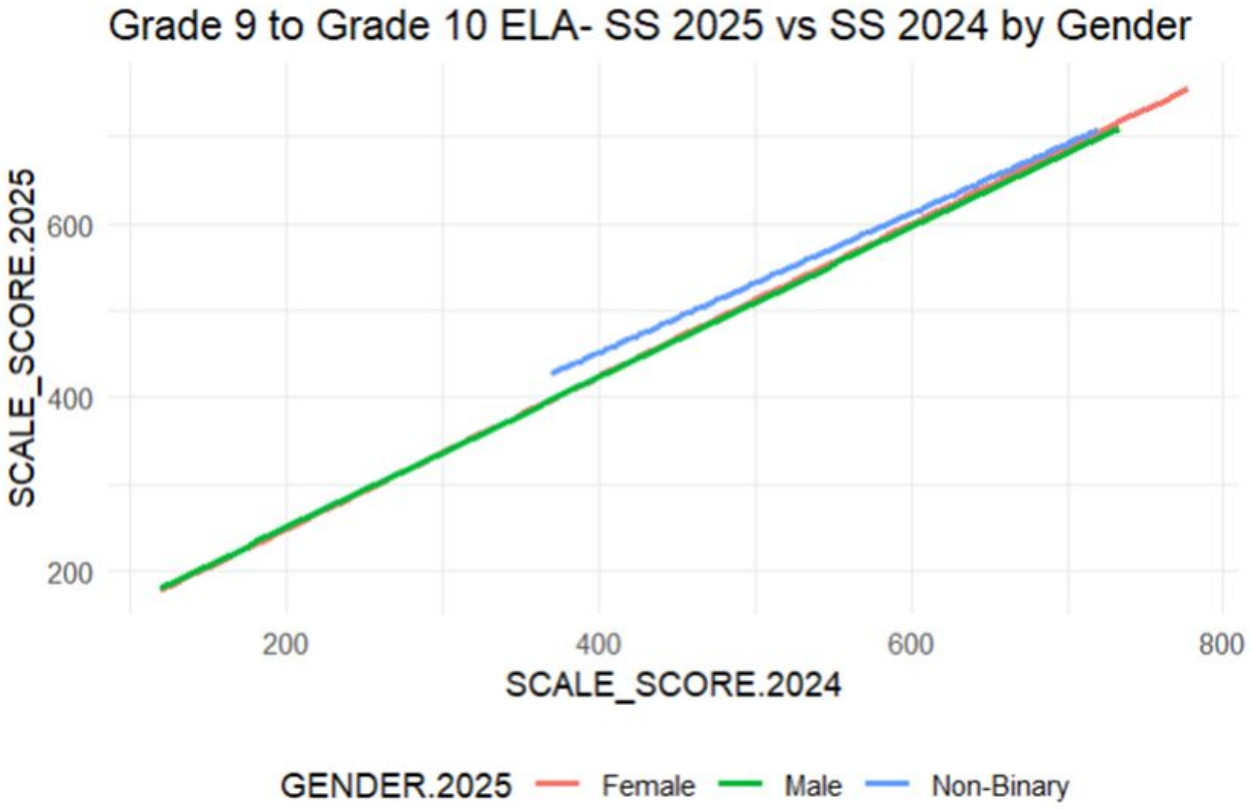
Historically higher performing male students show slightly larger scale score increases than female academic peers.

SGP to Scale Score Scatter plot for 2025: PSAT g9 Math by Gender



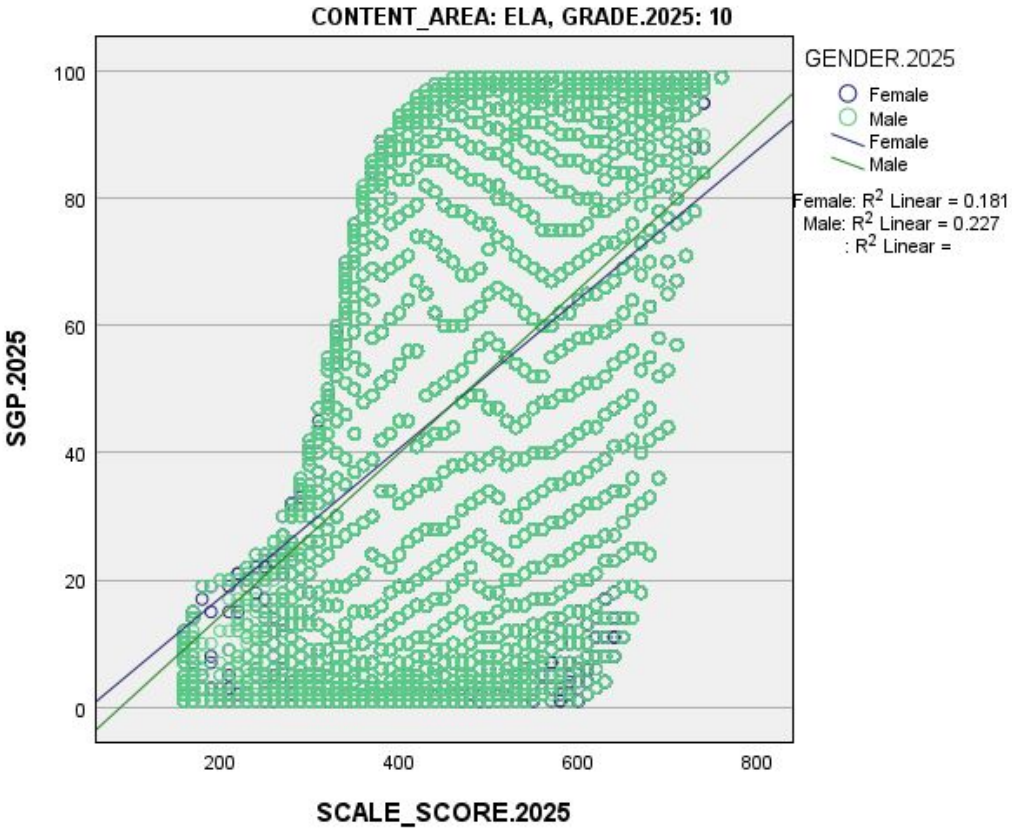
Exaggerating the trends seen in the scale score trends, historically lower performing male students show lower growth than female peers, but historically higher performing males show larger score increases than female peers.

Scale Score Trends from 2024 to 2025: PSAT g9 ELA to PSAT g10 RW by Gender



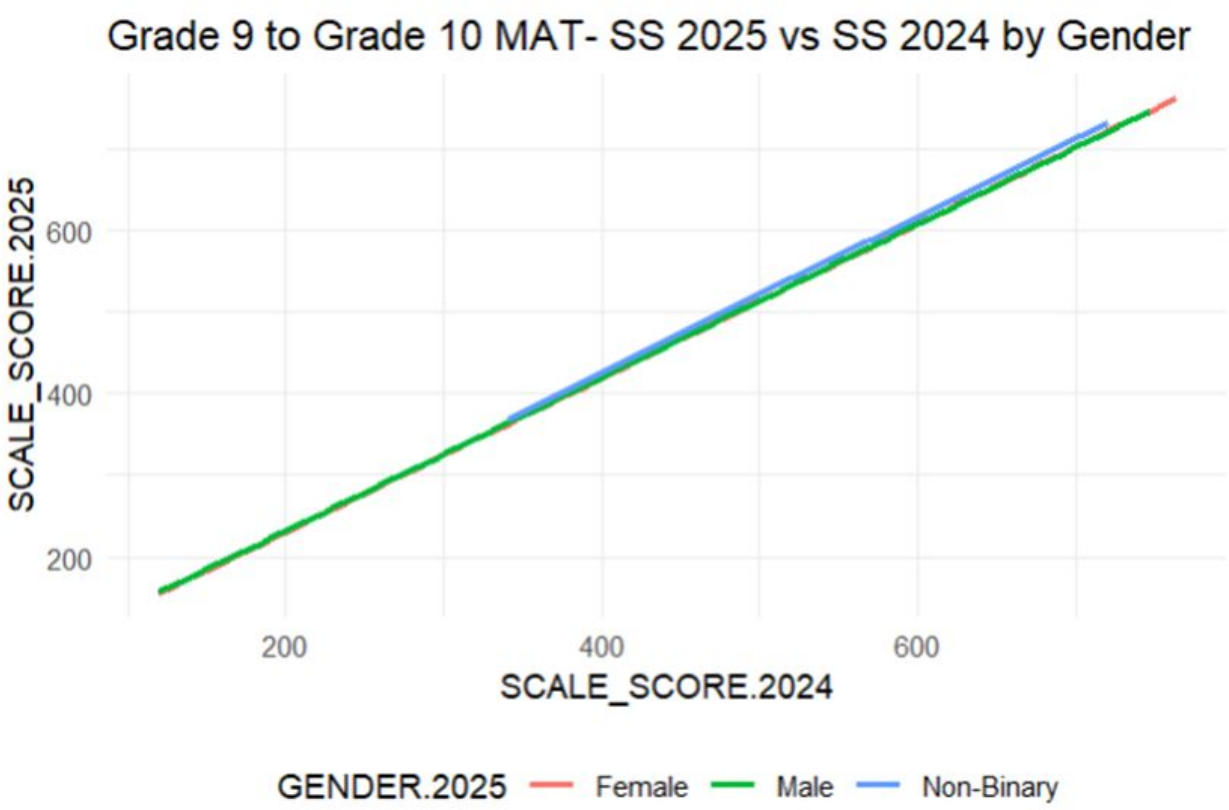
Nearly identical scale score increases for both male and female students.

SGP to Scale Score Scatter plot for 2025: PSAT g10 ELA/RW by Gender



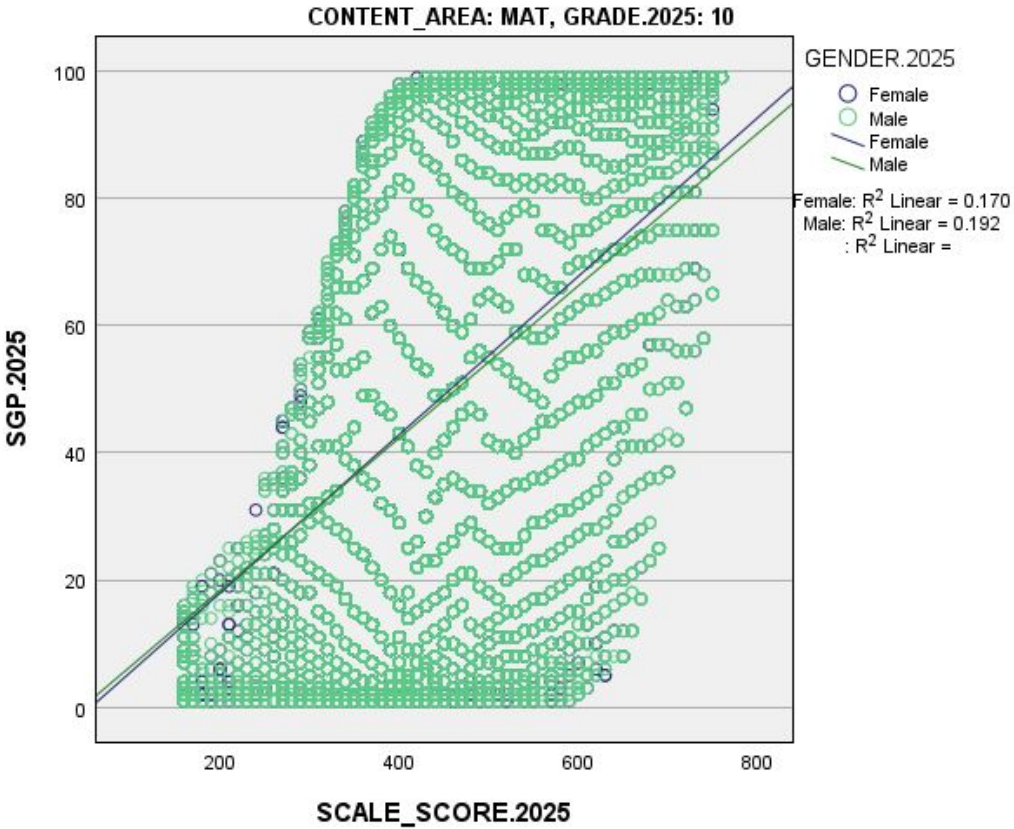
At odds with the scale score trends, historically lower performing male students show slightly lower growth than female peers, while historically higher performing males show larger score increases than female peers.

Scale Score Trends from 2024 to 2025: CMAS g8 to PSAT g9 Math by Gender



Nearly identical scale score increases for both male and female students.

SGP to Scale Score Scatter plot for 2024: PSAT g9 Math by Gender



Historically higher performing females show slightly higher SGPs than male peers.

Student-level Between-Year Scale Score Correlations

		2023 to 2024		2024 to 2025	
		N	Corr	N	Corr
ELA/RW	CMAS g3 to CMAS g4	51458	0.827	50410	0.829
	CMAS g4 to CMAS g5	51182	0.836	51296	0.826
	CMAS g5 to CMAS g6	50277	0.816	49845	0.832
	CMAS g6 to CMAS g7	48711	0.831	48070	0.834
	CMAS g7 to CMAS g8	45283	0.836	45237	0.844
	CMAS g8 to PSAT g9	44770	0.782	43872	0.787
	PSAT g9 to PSAT g10	50253	0.842	49656	0.863
	PSAT g10 to PSAT g11	49466	0.866	50350	0.870
Math	CMAS g3 to CMAS g4	53016	0.851	52354	0.863
	CMAS g4 to CMAS g5	52313	0.860	52709	0.865
	CMAS g5 to CMAS g6	50503	0.842	50433	0.858
	CMAS g6 to CMAS g7	48937	0.853	48548	0.853
	CMAS g7 to CMAS g8	45442	0.848	45825	0.852
	CMAS g8 to PSAT g9	44993	0.835	44162	0.842
	PSAT g9 to PSAT g10	50253	0.841	49656	0.876
	PSAT g10 to PSAT g11	49466	0.819	50350	0.890

The correlation between CMAS g8 ELA and PSAT g9 RW is the only one that falls below 0.8. This was the bar that the TAP used in 2018 to determine the ELA and EBRW constructs were not aligned enough to warrant calculating growth.

TAP Recommendation

Does the TAP recommend continuing to suspend calculating growth from CMAS g8 ELA to PSAT g9 Reading & Writing?

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: November 20, 2025