Indicator Scoring Analysis

Table of contents

Indicator Scoring Analysis

1 Executive Summary

1.1 Overview

The school grading system is designed to inform and empower parents, community members, and school leaders. An ideal system provides clear and actionable data, helping everyone work toward shared educational goals. The system must communicate accurately, minimize confusion, and remain consistent over time to be truly effective.

The current system operates admirably within the constraints of the statute and the available data. Many stakeholders across Arizona have invested significant time and effort to identify meaningful and agreed-upon metrics. However, four mathematical issues within the system undermine its effectiveness: the results are difficult to interpret and may not be sending the intended signals to stakeholders.

1.2 Key Findings: Four Mathematical Issues

- **Balance**: When combining data with very different distributions—like some being symmetric and normal while others are highly skewed—it can distort the results, as each contributes unevenly to the overall outcome.

- **Scaling**: When data from different sources or measures have varying ranges and impacts, combining them without standardizing can distort results.

- **Discretization Error**: Converting numerical data into categories (e.g. Low, Medium, High) too early can blur meaningful differences, especially when similar values are split into separate groups, leading to data loss and less accurate interpretations.

- **Sampling Error**: This occurs when the sample used to draw conclusions doesn't accurately represent the larger population, leading to potentially misleading or unreliable results.

1.3 Agile Solution Approach

Imagine designing the ideal vehicle: fast, safe, and comfortable. In a traditional process, one might spend months or even years refining every detail—from the engine components to the interior design—before revealing the finished product. If this final design does not meet users' needs, the discrepancy only becomes apparent after significant time and resources have already been invested. An agile methodology inverts this process. Instead of waiting until every aspect is perfected, development begins with something simple—a "skateboard," so to speak—which, though far from the ultimate vision, still provides immediate functionality. Users can test this initial version, offer feedback on what works or what requires improvement, and each iteration addresses these insights.

Progressing from the "skateboard" to a "scooter," then to a "bicycle," and subsequently to a "motorcycle," developers maintain user involvement throughout each stage. Over time, designers discover whether the handle design is comfortable, whether a seat is necessary, and whether additional storage capacity is required. By continuously releasing these functional prototypes, they gather frequent feedback and refine the product accordingly. Thus, by the time the final "car" is delivered, it reflects user requirements rather than assumptions fixed at the start.

Applying this approach to indicator scoring, the Accountability Unit intends to first release "skateboard" business rules in the summer: a basic, initial version of the Indicator Scoring model. By sharing these preliminary results with the Technical Advisory Committees and the broader field, immediate feedback can be gathered, and each iteration refined. Over time, the model will evolve to become more comprehensive and aligned with stakeholder needs, ultimately yielding a final approach that accurately and meaningfully captures school performance.

2 Explanation of Mathematical Issues

2.1 Balance

- **Definition:** When we combine data with very different distributions—some symmetric and "normal," others heavily skewed—it can create distorted results. Essentially, one measure can outweigh the rest simply because of how its values are spread out, rather than due to genuine differences in performance.

- **Example**: Imagine a seesaw with a single heavy book on one side and several smaller books on the other. This delicate arrangement stays balanced only if no book is removed or replaced. If you take away one small book or swap it for a heavier volume, the entire seesaw tips. A better design would ensure each book lies flat and shares the load evenly in the middle—so removing or replacing one wouldn't disrupt the whole system.

- **Why It Matters**: In our current accountability model, certain indicators are like that heavy book: almost no schools earn full points on those metrics. Other indicators, where almost everyone earns full points, act as "counterweights." This setup forces the system to rely on everything remaining exactly in place. If a school lacks data for one indicator (e.g., because it's new and has no graduates), or if we adjust how a single measure works, the whole system can become unbalanced. Schools that differ from the typical profile—such as small, rural, or homogeneous campuses—often feel this effect more acutely. Additionally, any attempt to improve one indicator affects the entire model, making meaningful changes complicated.

- **Recommendations**:

    - Ensure each indicator is individually "balanced" before combining them—for instance, by removing extreme outliers or using z-scores.

    - Consider setting clear proportions or cut points so one measure doesn't overshadow the rest.

2.2 Scaling

- **Definition**: When data from different measures vary widely in both range and impact, combining them without a standard approach can lead to distorted outcomes. Essentially, one measure might inadvertently carry far more weight—or far less—than intended, obscuring real performance differences.

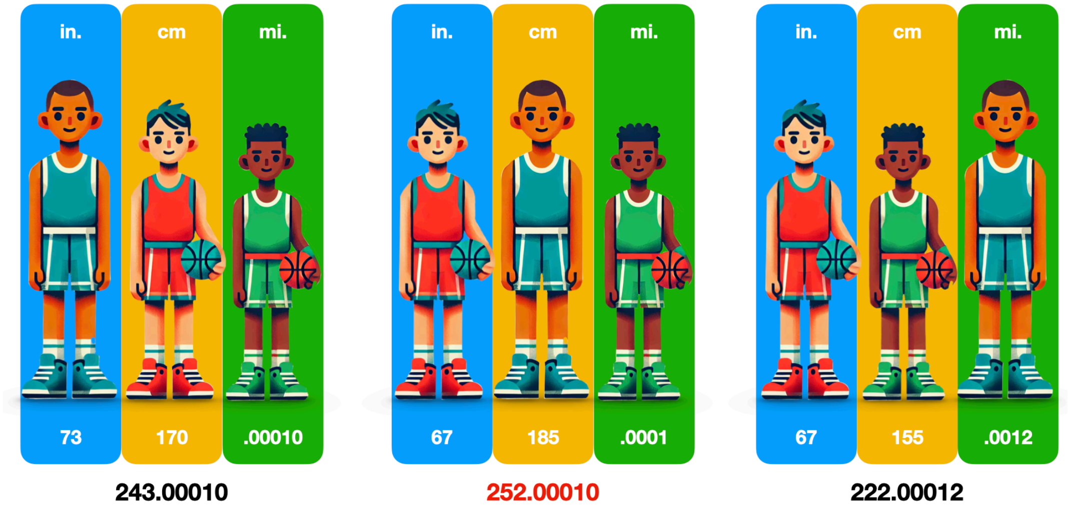
- **Example**: Imagine ranking youth basketball teams by measuring each player's height in inches, centimeters, and even miles—then simply adding all those measurements together. A final score of 152.4001 points would result from a 7-foot player measured in miles contributing only 0.001 "points," whereas a 5-foot player measured in centimeters adds 152.4 points. The final total ends up reflecting the metrics used rather than a true comparison of the players' heights.

- **Why It Matters**: In a school accountability context, our goal is to show how well a school meets the needs of all students. If one indicator covers a broad range of performance while others are more narrowly distributed, schools may be rewarded or penalized based on which indicators they perform well in, rather than their overall effectiveness. When improvements in one area barely move the needle, educators receive weak or misleading feedback—and may feel pressured to chase points rather than focus on what students actually need. Gradual, meaningful changes deserve to be captured by the system, but without proper scaling, even sizable improvements might not register.

- **Recommendations**:

  - Use Transparent Scaling Methods: Consider techniques that normalize different distributions, while still being easy for stakeholders to interpret.

- Respond to Incremental Growth: Design the score ranges so that small, steady improvements are visible and rewarded, preventing a “go big or get nothing” dynamic.
- Adjust Points Proportionally: Ensure that being in the top 10% of any indicator is worth at least 90% of its possible points, so each metric carries appropriate weight.



2.3 Discretization Error

- **Definition**: Converting numerical data into categories (e.g., Low, Medium, High) too early can blur important distinctions. Similar values may end up in separate categories, while large differences can be hidden, leading to data loss and less accurate conclusions.
- **Example**: Consider a bank that decides to categorize deposits in multiples of 20 for ATM convenience. Anything under $10 becomes 0 (“small”), anything over $75 becomes $100 (“large”), and anything in between becomes $20 (“medium”). While many customers might find their balance roughly evens out over time, edge cases can lead to extreme discrepancies. For instance, five deposits of $74.99 would be rounded to five “medium” deposits for a total of $100, while five deposits of $75.01 would be categorized as five “large” deposits, ending up at $500. A better real-world approach is to add up all deposits exactly, then convert to convenient withdrawal categories only at the end. This minimizes the overall distortion caused by any single categorization step.
- **Why It Matters**: In our current accountability model, we sometimes create categories too early, then aggregate those categories again—compounding the data loss at each step. By the final stage, crucial information about performance has been stripped away. Preserving the original, more detailed data as long as possible

allows for a more accurate analysis before any final categorization occurs. Every time we move from continuous numbers to discrete groups, we inevitably lose detail, which can misrepresent real performance trends or improvements.

- **Recommendations**:
  - Analyze Before Categorizing: Delay classification until after we've extracted all necessary insights from the raw data.
  - Reduce Multiple Rounds of Grouping: Each additional conversion step can magnify data loss.
  - Ensure Meaningful Breakpoints: When categories are needed, establish cut points that represent genuine differences in performance rather than arbitrary thresholds.

2.4 Sampling Error

- **Definition**: Sampling error occurs when the group used to draw conclusions does not accurately represent the broader population, potentially leading to misleading or unreliable results.

- **Example**: Imagine a statewide contest to find Arizona's "best" farmer, where each contestant picks a random bag of seeds. Someone who picks a bag of watermelon seeds will likely outproduce a contestant who picks strawberries—making luck, rather than skill, the biggest factor. A fairer approach would give every farmer an identical, mixed bag of seeds. In that scenario, the harvest outcome would be determined more by the farmers' abilities and less by chance.

- **Why It Matters**: School enrollment isn't designed to create a perfectly representative sample of students from the entire state. Each school serves a unique community with distinct needs, so grouping students by school can introduce sampling biases. A school might look especially strong or weak in certain measures simply because of the particular student population it serves, not because of the school's actual practices. Recognizing and accounting for these biases helps ensure that accountability results are interpreted fairly.

- **Recommendations**:
  - Identify Potential Biases: Determine where sampling issues may arise, such as in specialized programs or regional demographics.
  - Control for Biases Where Possible: Consider statistical or policy-based controls (e.g., subgroup analyses) to level the playing field.
  - Avoid Overgeneralization: Design measures with the recognition that differences observed among schools may reflect sampling variations rather than actual performance gaps.

A-F System and Indicator Analysis

3 Summative System

3.1 Issues

- Balance
- Scaling

3.2 Description

- Some indicators within the current summative grading model are challenging enough that almost no schools earn full points, while others are comparatively more attainable and yield near-perfect scores for most schools. As a result, if a school is ineligible for one of the indicators, its overall balance shifts unexpectedly. Moreover, even minor adjustments to a single indicator can disrupt the equilibrium for all schools and necessitate changes to the summative cut scores.

- Because the indicators in the accountability model operate on different scales, high performance in one area may yield significantly more points than equivalent performance in another. This discrepancy is evident in the K-8 model: while the top 10% of schools earn between 87% and 100% of the available points in Proficiency, schools in the top 10% for every other indicator receive full points. These variations frequently cause confusion about why the threshold for an "A" might be 84% instead of 90%. In practice, it is easier to achieve 100% of the points in some indicators than in others, and these discrepancies affect the overall summative scores.

- When all points are combined and averaged, the model often "requires" near-perfect performance in indicators that typically yield higher scores—such as Growth and Acceleration/Readiness—to counterbalance indicators like Proficiency. Bonus points introduce an additional scaling challenge, since nearly twice as many bonus points are available in some high school models compared to K-8 models. These factors underscore the need for a more consistent approach to scaling across all indicators.

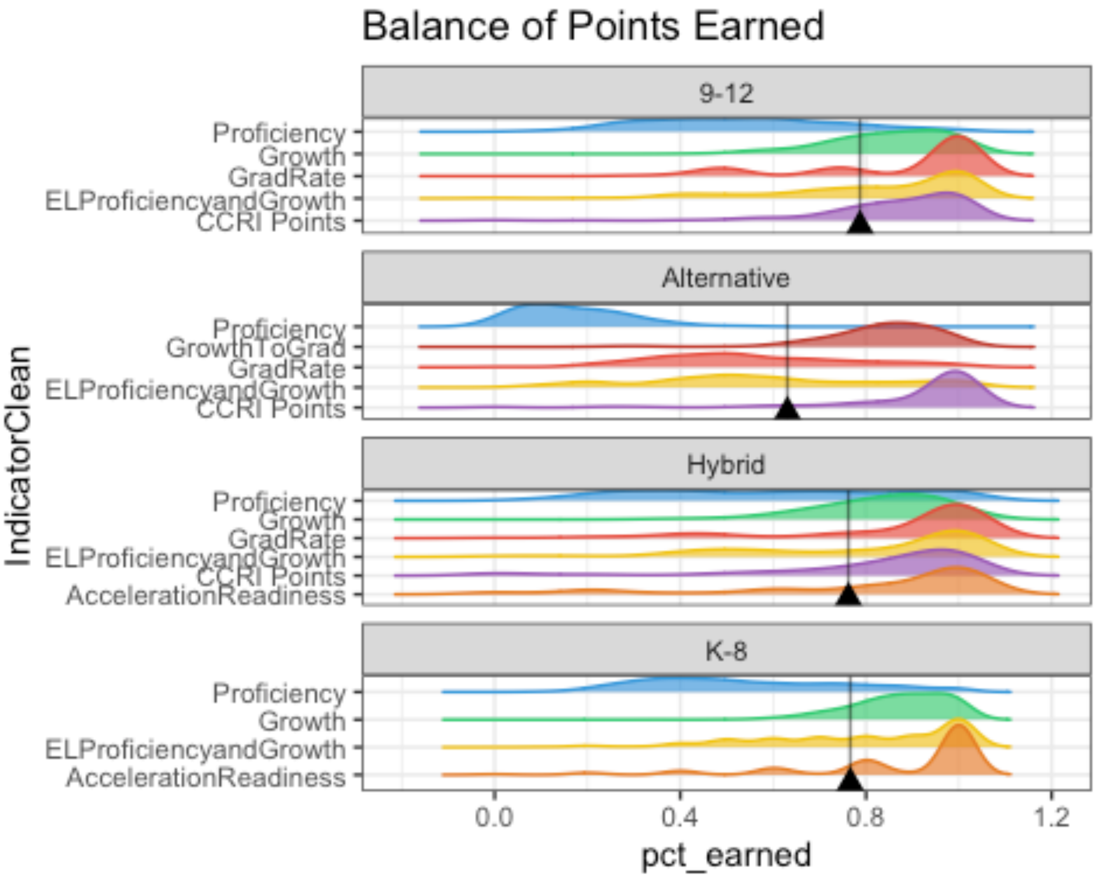
3.3 Recommendations

- **Establish Uniform Point Allocations**: Explore methods that produce consistent outcomes for schools demonstrating similar performance. For instance, ensure that "A" schools receive between 90% and 100% of the available points in an indicator, "B" schools earn between 80% and 89%, and so on.

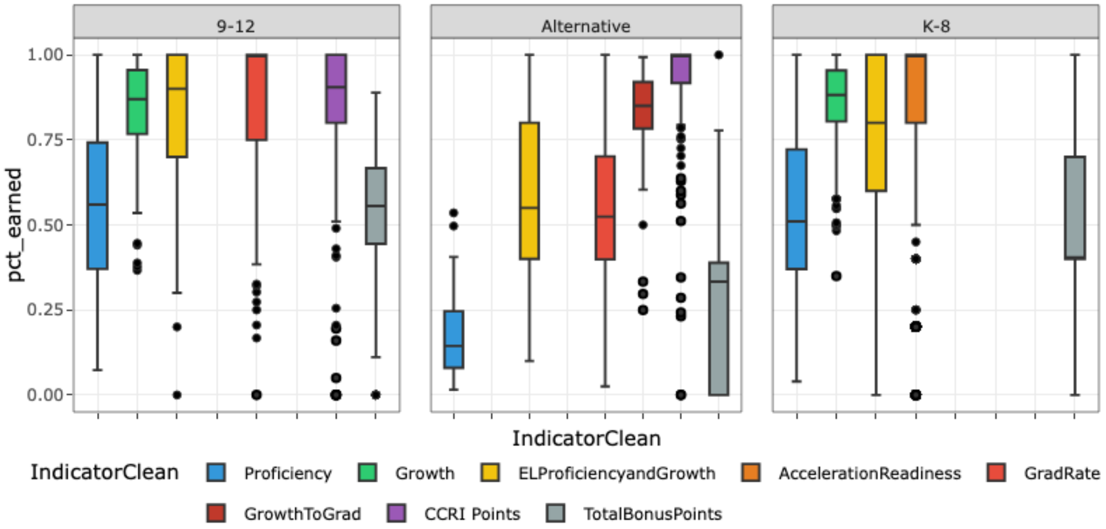
- **Use an Additive Approach for Summative Points**: Using a GPA-style calculation reintroduces scaling discrepancies, rendering lower-weighted indicators inconsequential. An additive model helps maintain each indicator's intended influence on a school's overall grade.

3.4 Visuals

The plot below demonstrates how the current system's balance mechanisms function. The densest portionin each distribution indicates the percentage of points earned by the majority of schools. Indicators in which most schools earn relatively few points are offset by indicators where schools tend to earn a higher percentage of points. The vertical lines represent the average percentage of points earned across all indicators, serving as a rough "balancing point."



This plot illustrates scaling concerns by showing the actual range of points schools earn on each indicator—despite the theoretical 0–100% span. Some indicators, such as the EL measure in the Alternative model, display a broad distribution from 10% to 100% of possible points (a 90-point range). Others, like the CCRI indicator, are confined to roughly 80%–100% (a 20-point range). Consequently, each percentage increase represents a different portion of the total available points, yet in the final summative calculation, each point increment is treated equally.



4 Proficiency

4.1 Issues

- Sampling Error
- Balance (minor)

4.2 Description

- The primary concern with the Proficiency indicator is sampling error. Because students are not randomly assigned to schools based on their proficiency levels, the indicator partly reflects variations introduced by this non-random distribution rather than solely capturing school-level actions.
- Minor balance issues do arise from the point intervals (0, 0.6, 1.0, and 1.3) and the possibility of earning more than full points, which slightly skew the distribution. However, the impact is negligible. Adjusting these intervals may be beneficial for communication rather than for purely statistical reasons. The current weights, while convenient for calculation, obscure performance levels. Although performance levels are on a 1–4 scale with equal increments, an average proficiency score of 0.58 is not readily interpretable in terms of actual proficiency. Since the system's overarching goal is to provide actionable insight rather than mere incentives, it may be worthwhile to consider assigning points in a way that clarifies meaning. A 1–4 performance scale has a nearly symmetrical distribution that aligns well with the existing level assignments.

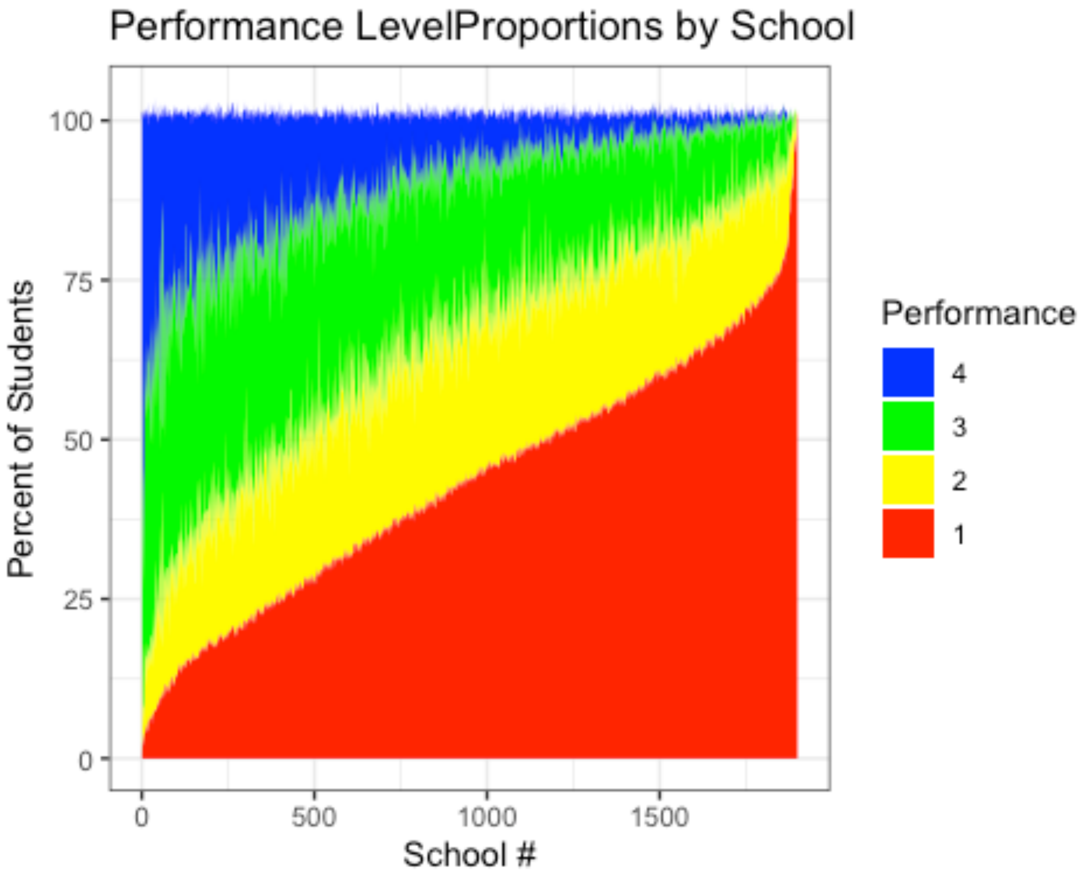
4.3 Recommendations

- **Control for Sampling Error**: Employ methods that directly address sampling issues, such as statistical adjustments.

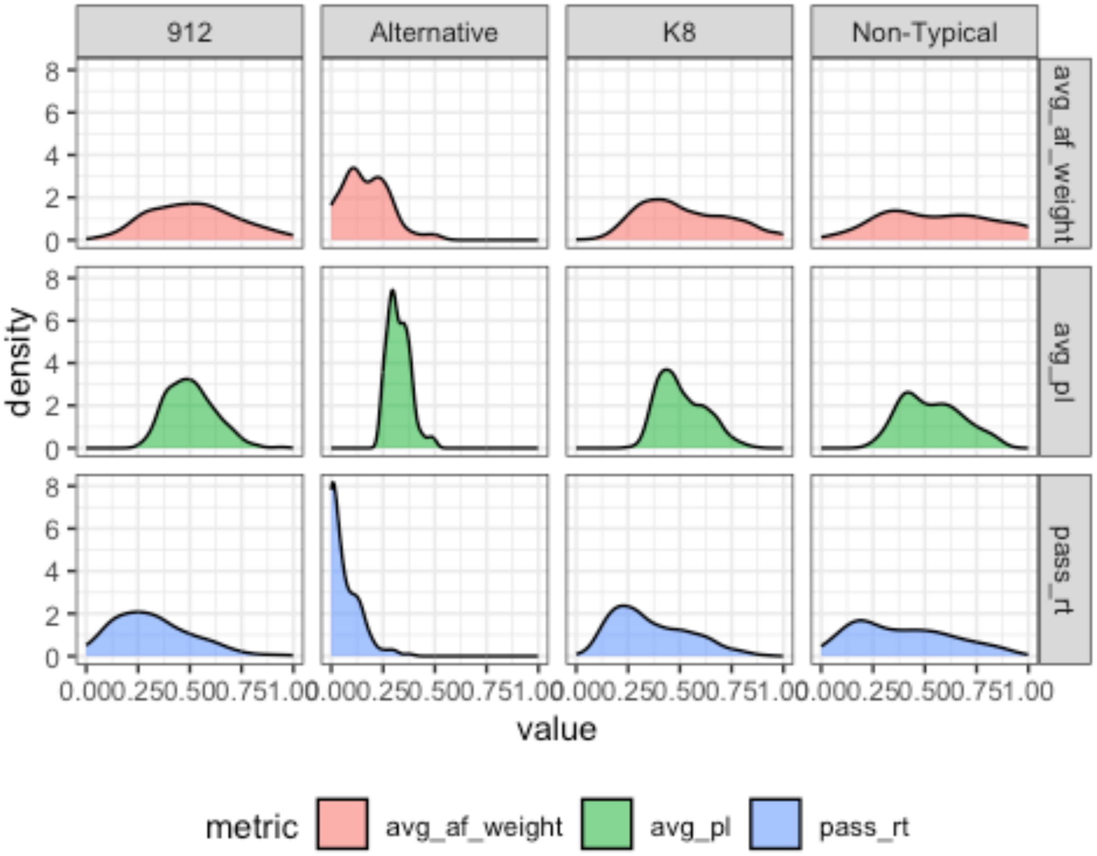
- **Combine Proficiency and Growth**: Consider merging Proficiency with Growth into a single indicator, similar to the current structure in other areas (e.g., EL Proficiency and Growth), to reduce sampling variance and improve overall reliability.

4.4 Visuals

This chart depicts the proportion of students entering each school by their prior-year assessment performance level. In a purely random assignment, we would expect to see uniform "bands" across the entire chart. However, because student enrollment is not random, schools begin each academic year with distinctly different achievement profiles.



This chart depicts distribution outcomes for three metrics: (1) Average A-F Weights, derived from the current 0, 0.6, 1.0, and 1.3 scoring system; (2) Average Performance Level; and (3) Pass Rate. Both Average A-F Weights and Average Performance Level are relatively balanced and exhibit approximately normal distributions. Pass Rate also follows a roughly normal pattern but is slightly right-skewed, suggesting that it may benefit from a statistical adjustment—such as treating outliers—to achieve greater overall balance.



5 Growth

5.1 Issues

- Balance
- Discretization Error
- Sampling Error

5.2 Description

- **Balance**: The principal balance concern in the Growth indicator stems from the ability to earn over 100% of the points, even though the maximum award is capped at 50. This design creates a markedly left-skewed distribution. While weighting prior-year performance is intended to reward strong growth among students who were previously low-proficiency, it can inadvertently disadvantage schools serving these populations. Because most schools earn or approach the full 50 Growth points, overall variability is limited. Approximately half of schools earn 45 points or more, 80% earn at least 40, and 94% earn 35 or more. For schools that do not reach near-full Growth points, the deficit can be significant, effectively making maximum Growth a prerequisite for securing a grade above "C."

- **Sampling Error**: Sampling error further amplifies Growth imbalances by tying a school's Growth outcome to factors beyond its control, such as the composition of its incoming students. In schools where more students begin at minimal proficiency (and are therefore eligible to earn double Growth points), the total number of available Growth points becomes disproportionately large. This weighting was meant to offset Proficiency's influence but has instead introduced imbalance into the Growth measure itself.

  A more effective approach would ensure each indicator is internally balanced, rather than using a left-skewed measure to compensate for a right-skewed one. Ideally, Growth should reflect instructional impact rather than simply adjusting for variations in incoming proficiency. However, the current transformation of Growth scores by proficiency status replaces a direct measure of school-level effort with a system that is negatively correlated to a school's existing proficiency profile—introducing additional sampling bias rather than offering clear feedback.

- Discretization Error: Currently, student growth is categorized into Low (0), Average (1), or High (2), allowing prior-year weighting to be applied. While this approach aids planning at the student level, it discards valuable nuance. A more refined method would calculate an average Student Growth Percentile (SGP) on a continuous 0–100 scale, then assign categories (e.g., A, B, C, D, or F) only after analyzing these raw data.

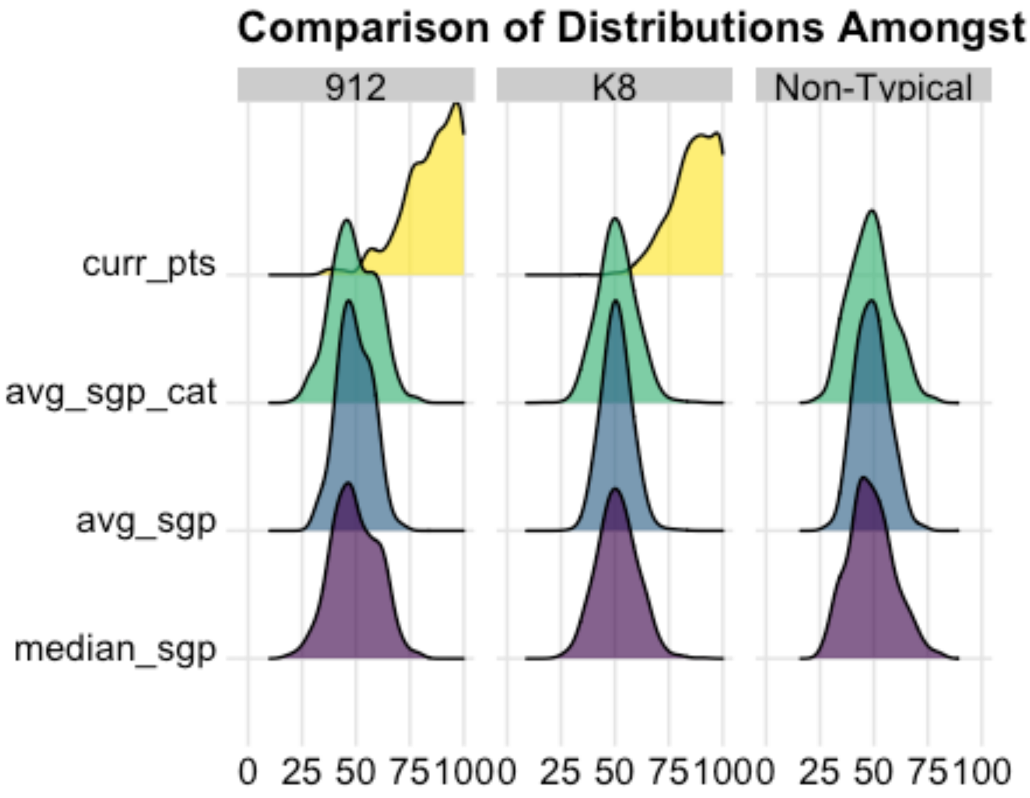
  Although large schools may experience less volatility due to sample size, smaller schools—or those with fewer students eligible for Growth—can see disproportionate swings. For example, if an elementary school's students all scored at the 33rd growth percentile, the school would earn zero Growth points; yet at the 34th percentile, it would earn a full 50. Although real-world differences are typically less stark, schools with many students near a categorical boundary can experience sudden, significant changes in Growth scores.

5.3 Recommendations

- **Use a Continuous, Unweighted Growth Calculation**: Preserve data during aggregation to create a balanced distribution, apply categories only after examining the full distribution of raw growth values.

5.4 Visuals

This graph compares four ways of calculating Growth. The top row (curr_pts) shows the current points-based method, which has noticeably skewed distributions—indicating that many schools cluster near the upper limit. Moving down to avg_sgp_cat shows some improvement, but there is still moderate skew. By contrast, both avg_sgp and median_sgp (the third and fourth rows) produce more symmetrical, centered distributions across all school types. In other words, these continuous growth measures avoid the extreme clustering found in the current approach and provide a more balanced indicator.



6 English Learners

6.1 Issues

- Balance

6.2 Description

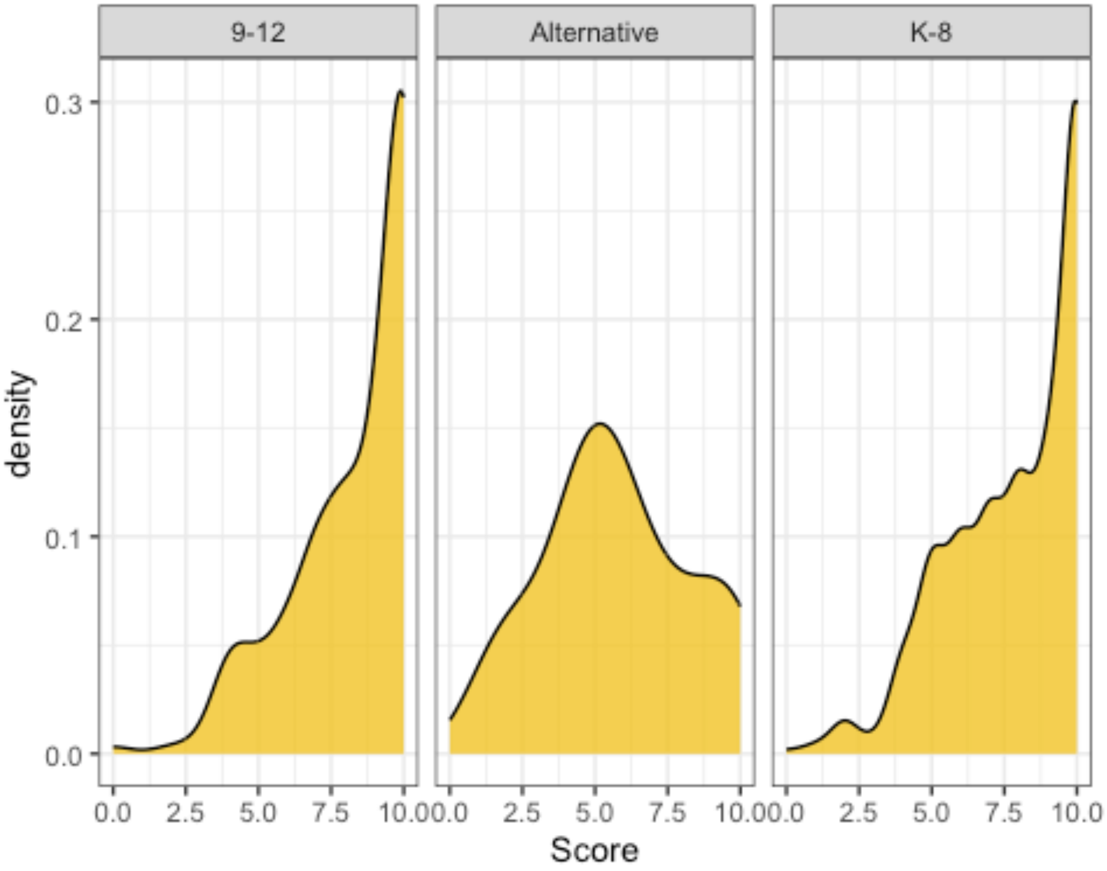
- Similar to Growth, the EL Indicator's final point distribution tends to be skewed. In an attempt to mitigate this imbalance, the system first removes outliers. However, the distribution is then shifted again by granting full points to any school that meets or exceeds the adjusted average. This process effectively "tips" the distribution to the right, where numerous schools cluster at or near the maximum score. Such a setup makes year-to-year results unpredictable for school leaders; annual fluctuations in state averages can alter whether a given school surpassed the threshold to earn points.

6.3 Recommendations

- **Preserve the Full Distribution**: Use an approach that balances outliers without unintentionally clustering scores at the high end. For instance, explore techniques that recognize legitimate top performers but still maintain a more even spread across the full score range.

- **Establish Stable Targets**: Rather than recalculating the threshold each year based on state averages, consider adopting reclassification and growth rate goals that remain consistent. School leaders could then direct their improvement strategies toward a fixed benchmark rather than waiting to see how the statewide average shifts.

6.4 Visuals



7 Acceleration/Readiness

7.1 Issues

- Balance
- Discretization Error
- Sampling Error

7.2 Description

- The Acceleration Readiness indicator faces several interrelated challenges of Balance, Discretization Error, and Sampling Error. Like Growth and EL, most schools earn near-maximum points, effectively making “full points” the floor. Because schools can accumulate up to 20 points but are ultimately capped at 10, the final distribution is heavily left-skewed. Within the indicator itself, certain components also exhibit balance issues: for instance, subgroup improvement awards 6 points out of a possible 52 (skewing the distribution), whereas Chronic Absenteeism either grants full points or none (producing a more normal distribution).

- Discretization and sampling errors arise because the system treats minute changes in subgroup proficiency (as small as 0.01) in a binary fashion, essentially a "coin toss". If a subgroup's proficiency shifts from 0.50 to 0.51, the school receives full points, but moving from 0.50 to 0.499 yields no points at all. Larger, more diverse schools typically have more "coin flips" (i.e., multiple subgroups or grade levels), thus increasing their chances of earning points by pure chance alone. Smaller schools or subgroups, however, may experience disproportionate volatility due to small shifts in enrollment and fewer attempts at improvement. Since year-to-year comparisons do not necessarily track the same students but rather the same demographic subgroup, sampling variation can further distort results.

7.3 Recommendations

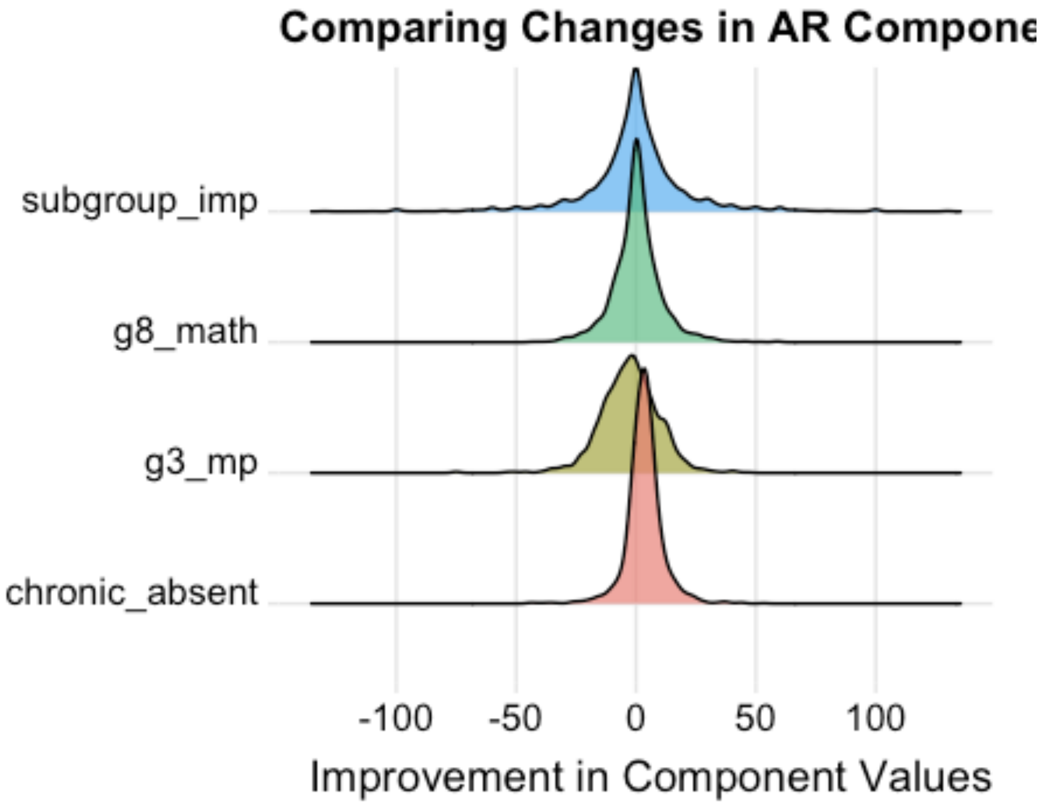
- **Limit or Remove Over-Allocation**: Reassess the practice of allowing more than the final cap of points before summing to 10. Measuring a school across all eligible components can mitigate left-skewed distributions and help maintain a clearer picture of performance.

- **Reduce Binary Thresholds**: Move away from strict all-or-nothing scoring (e.g., a 0.01 difference). Instead, consider a tiered or continuous scale that captures partial improvements without swinging from zero to full points in a single step.

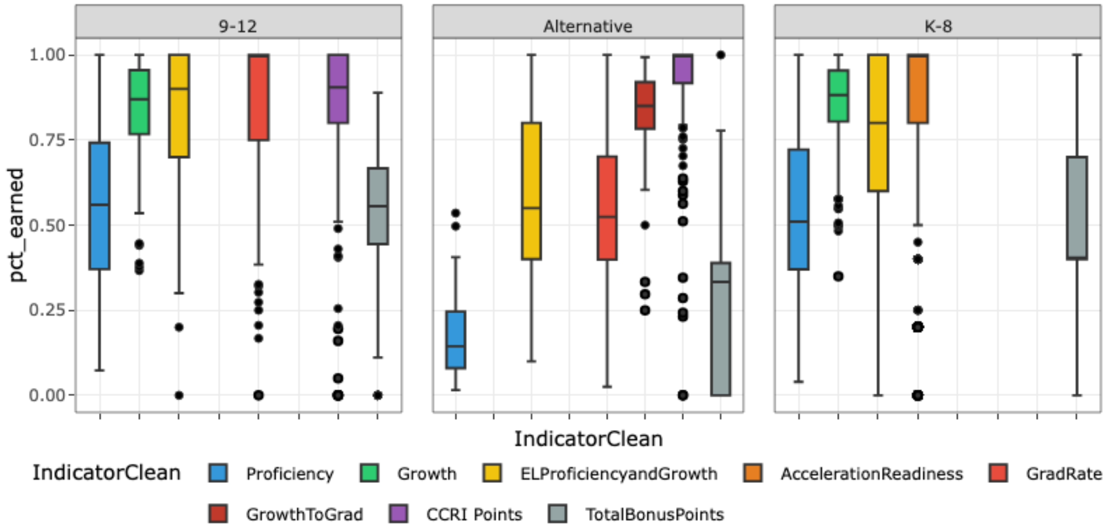
- **Track the Same Students Where Possible**: Consider an approach to subgroup comparisons that uses SGP data. Evaluating improvements among the same students year over year can reduce the impact of random variation and more accurately reflect a school's actual efforts.

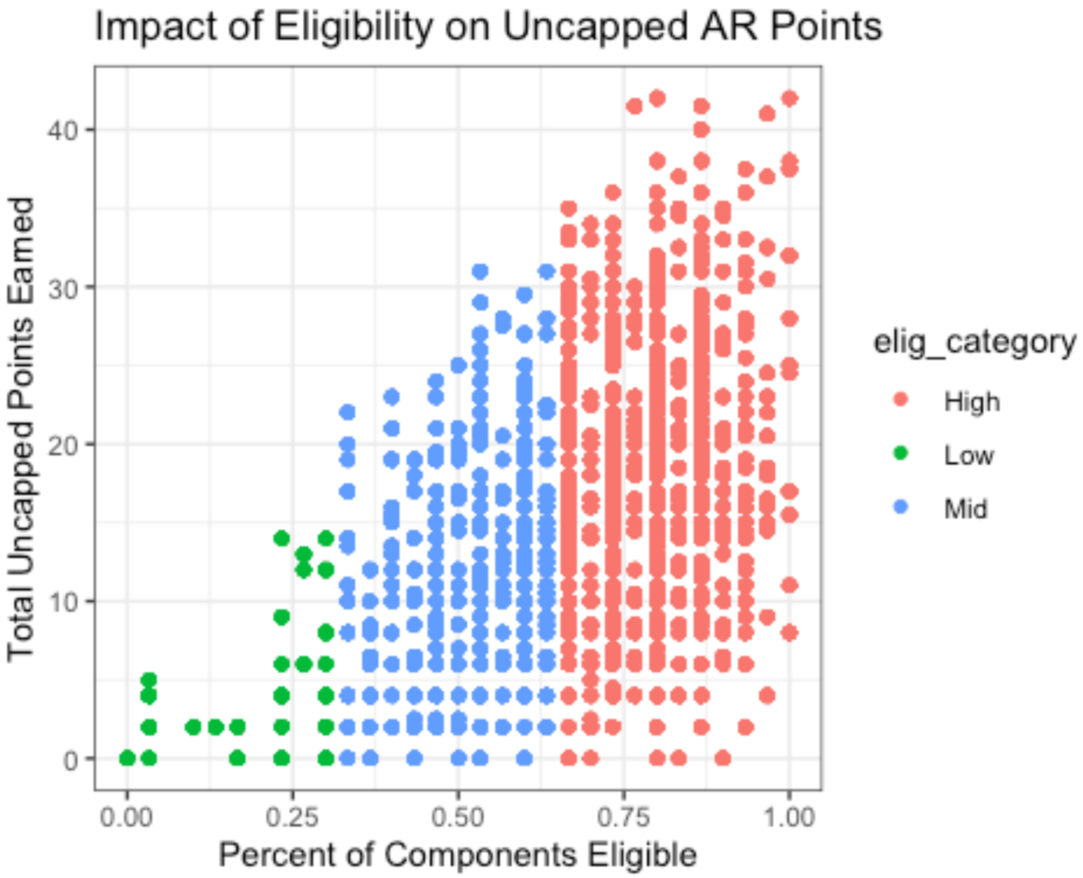
7.4 Visuals

The visuals illustrate the balance issues within the Acceleration/Readiness components. The first plot highlights that while the raw calculations for individual components are balanced, the process of awarding points introduces imbalance. This imbalance arises from discretization and the potential for schools to earn more points than the maximum possible.



includes change values for each individual subgroup and subject





8 Graduation

8.1 Issues

- Balance
- Discretization Error

8.2 Description

- The Graduation Indicator faces challenges related to Discretization Error and Balance. For improvement scoring, a school's graduation rate can either lose more than 2%, remain within ±2%, or gain at least 2%, yielding 0, 5, or 10 points respectively. This all-or-nothing approach exaggerates small differences in performance, resulting in significant point swings. Furthermore, assigning equal weight (10 points each) to Graduation Rate and Graduation Rate Improvement can produce counterintuitive outcomes. A school with a 15% graduation rate that improves by 3% might earn the same final points as a school with an 87% rate that drops by 3%, potentially receiving the same overall letter grade. Finally, the cap of 10 points out of a possible 12 leads to a slight skew in the distribution of graduation rate scores.

8.3 Recommendations

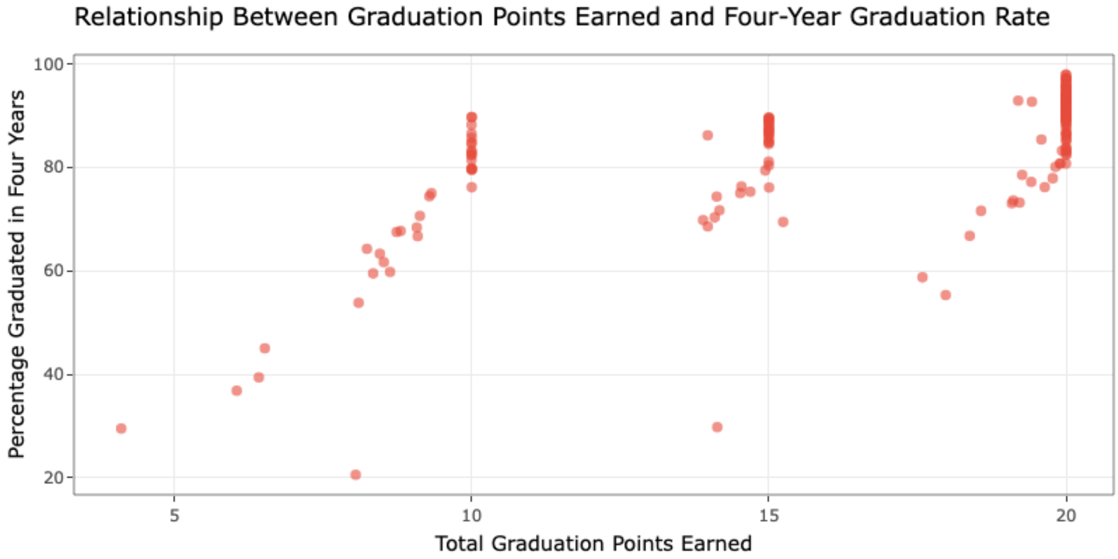
- **Adopt a More Nuanced Improvement Scale**: Move beyond the 0–5–10 point thresholds to a continuous or tiered scoring model that recognizes more incremental gains or losses in graduation rate.

- **Revisit Weighting Between Rate and Improvement**: Ensure that the weighting scheme reflects the relative importance of current graduation rates versus growth. Consider adjusting the points to avoid scenarios where drastically different rates yield identical final scores.

- **Reduce or Eliminate Point Caps**: Evaluate whether the 10-point cap (out of 12 possible) is necessary. Removing or adjusting it could mitigate skew and better reflect the actual range of performance.

8.4 Visuals

This scatterplot illustrates the relationship between the total graduation points earned by schools and their four-year graduation rates. Each point represents a school, with its graduation rate plotted against the total points earned for graduation-related indicators. As expected, there is a general association between higher graduation rates and higher graduation points. However, the variability in graduation rates within each point range is notable. For instance, schools with graduation rates as high as 89.7% may earn 10, 15, or 20 points, due to differences in points assigned for Graduation Rate Improvement.



9 CCRI

9.1 Issues

- Balance

9.2 Description

- The CCRI Indicator exhibits pronounced skew—particularly within the Alternative Model—where most schools accumulate near-maximum points. This creates a “light switch” effect, in which schools either achieve a near-perfect score or appear significantly underperforming. Consequently, the indicator may offer limited nuance in reflecting true differences among schools’ college and career readiness practices.

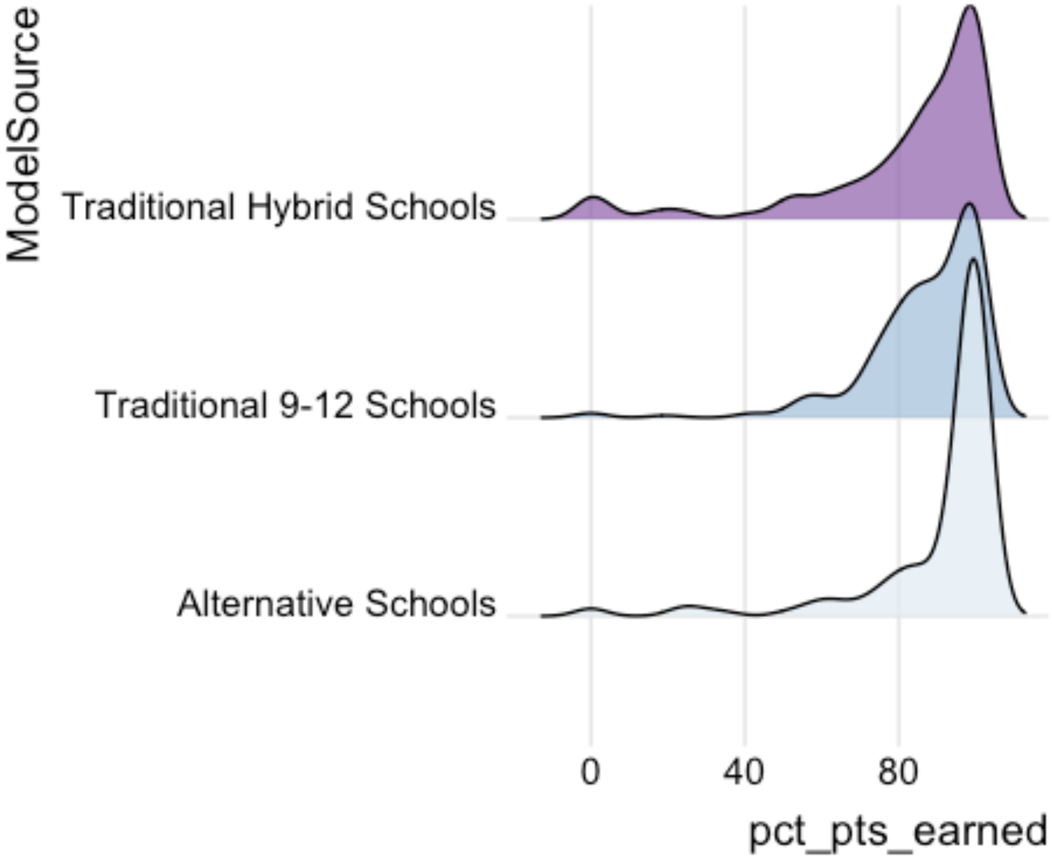
9.3 Recommendations

- **Revisit Traditional CCRI Point Allocations**: To address potential redundancies, consider combining or capping points awarded for components that measure similar constructs. For example, students currently earn points for ACT, SAT, and Accuplacer performance, which assess comparable skills. These measures could be grouped under a single category, such as “College Entrance Exams,” with a capped point allocation. Whether a student succeeds on one or all three exams, the total contribution to CCRI would remain the same.

- **Expand the Required Number of Alternative Components**: Currently, students in Alternative Schools can achieve a CCRI score of 35 by earning a single point in any area. It may be worth exploring whether the metric should differentiate between schools that help students earn points across multiple areas versus those focusing on a single area. However, this consideration must weigh potential benefits to students against the administrative burden it may impose on schools. Expanding requirements could encourage more comprehensive readiness practices but might also introduce unnecessary complexity.

9.4 Visuals

The visual illustrates the skewed distribution of points in the CCRI Indicator, where most schools cluster near the maximum, reflecting limited differentiation.



10 Growth to Graduation

10.1 Issues

- Balance

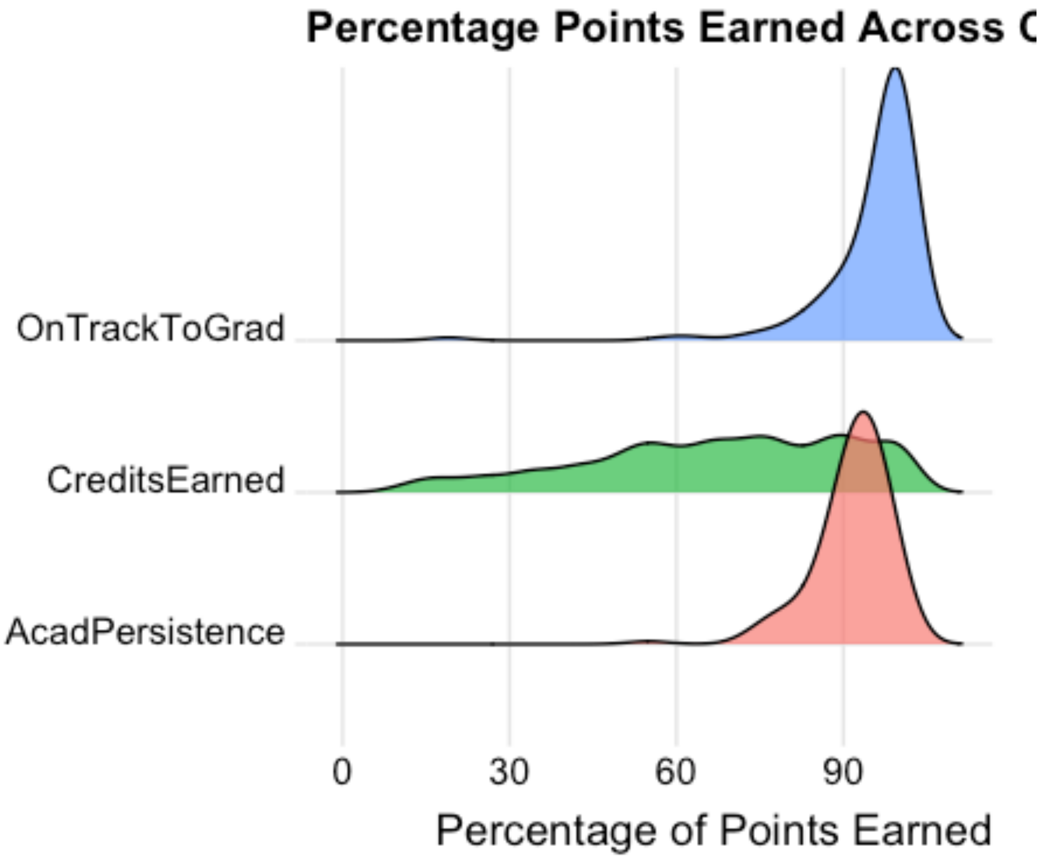
10.2 Description

- Discussion is needed to know whether the significant skew in the Persistance and On-Track-to-Graduate componenets is problematic, or is an actual reflection of the performance expectations. Currently, these two components signal that a school is underperforming if it is not achieving roughly 90% in these two measures. Unbalanced metrics is going to have less influence overall on summative grade. Here, credits earned will be the main factor that distinguishes schools' performance in the Growth to Graduation Indicator.

10.3 Recommendations

- Discuss to determine whether the current dynamics of the indicator are a feature or a problem in the Indicator.

10.4 Visuals


Percentage Points Earned Across C
OnTrackToGrad
CreditsEarned
AcadPersistence
0
30
60
90
Percentage of Points Earned