CADRE REPORT

Nominal and Effective Weights of Composite Accountability Ratings:

A Demonstration Using Colorado's School Performance Framework

By Benjamin R. Shear and Kaitlin Nath

August 2024

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

School of Education
UNIVERSITY OF COLORADO **BOULDER**

Nominal and Effective Weights of Composite Accountability Ratings:

A Demonstration Using Colorado's School Performance Framework

August 2024

Benjamin R. Shear
Kaitlin Nath

Center for Assessment, Design, Research and Evaluation (CADRE)
University of Colorado Boulder

Please send correspondence concerning this report to Benjamin Shear, benjamin.shear@colorado.edu. The authors acknowledge helpful feedback on earlier drafts of this report from Elena Diaz-Bilello, Erik Whitfield, Marie Huchton, and Dan Mangan.

Suggested citation: Shear, B.R., & Nath, K. (2024). *Nominal and effective weights of composite accountability ratings: A demonstration using Colorado's School Performance Framework*. Boulder, CO: The Center for Assessment, Design, Research and Evaluation (CADRE), University of Colorado Boulder.

Abstract

Federal law requires states to create school accountability ratings based on multiple factors including student academic achievement, academic growth, and progress in achieving English language proficiency. Many states meet these requirements by creating weighted composite scores based on indicators for each factor. When a composite score is constructed as a weighted sum of multiple indicators, the weights are called nominal weights because they represent the intended weights being given to each indicator. The actual influence of each indicator on the overall composite is referred to as the effective weight of that indicator. When the nominal and effective weights of an indicator differ, it can undermine the validity of interpretations of the composite scores. We review a widely used method for calculating effective weights and apply it to evaluate the alignment of nominal and effective weights for composite accountability scores produced in the Colorado School Performance Framework (SPF). We find the alignment of the nominal and effective weights varies across grade levels, with higher alignment for the elementary and middle school scores than for high school scores. We discuss factors that influence this alignment, such as the correlation and relative variances among the included indicators, and implications for designing school accountability composite scores.

Introduction

The Every Student Succeeds Act (ESSA) passed in 2015 requires states to create school accountability systems to "improve student academic achievement and school success." These school accountability systems are intended to help the state monitor student progress, identify schools requiring additional support, and ultimately improve educational opportunities for all students. ESSA specifies five factors that states must include in their systems to identify schools for comprehensive support and improvement: (1) student academic achievement in math and English Language Arts (ELA), (2) student growth in math and ELA (or an alternative), (3) progress in achieving English language proficiency, (4) graduation rates (for high schools), and (5) one additional measure of "school quality or student success" (SQSS). ESSA provides states flexibility in creating indicators for each factor, but states are required to construct an overall rating for each school that considers performance on all five indicators for all students and for each student subgroup.

States can meet these requirements by creating a weighted composite score from indicators of each factor, with weights assigned to align with policy priorities. Many states, for example, assign the largest weight to the growth indicator to reflect the belief that student growth provides a more valid indicator of school quality for accountability purposes than average achievement (Erwin et al., 2021). In Colorado the state assigns weights of 60% and 40% to indicators of student growth and achievement,[1] respectively, and similar approaches have been used in other states. Placing greater weight on growth indicators relative to achievement indicators is accompanied, at least implicitly, by the assumption that growth indicators should have a greater influence on school ratings and accountability decisions than achievement indicators. As we show in this report, however, the indicators assigned the

[1] These are weights used for identification in Colorado's state accountability system, used as the example in this report. The weights used to construct accountability indicators for Federal reporting under Colorado's recently approved ESSA plan differ slightly but still assign the highest weight to growth. See: https://www.cde.state.co.us/fedprograms/co-consolidatedstateplan-finalamended2023

greatest weight may not always have the greatest influence on overall composite ratings. This could pose a threat to the validity of composite accountability ratings.

This report proceeds as follows. First, we review indicator weighting and describe the difference between nominal and effective weights. We then introduce the Colorado School Performance Framework (SPF) school accountability rating system, which we use to demonstrate interpretation and comparison of nominal and effective weights in a school accountability system. The Colorado SPF ratings are used by the state to identify successful schools for recognition and schools in need of greater support. Although the indicators and weights used in the SPF differ from those used in Colorado's Federal ESSA identification process, the primary indicators and overall approach are similar. We end with summary comments and discussion.

Nominal and Effective Weights

This section describes the distinction between nominal and effective weights, and then presents one widely used method to calculate effective weights. When a composite score is constructed as a weighted sum of multiple indicators, the weights are called "nominal weights" because they represent the intended weights being given to each indicator (Wang & Stanley, 1970). Nominal weights are most often set judgmentally to align with system goals by assigning larger weights to indicators deemed more important (Baldwin, 2015; Wang & Stanley, 1970).

Under ESSA, for example, all indicators must be assigned "substantial weight" with "much greater weight" afforded to indicators of the first four factors relative to the weight assigned to the SQSS indicator. These requirements have been applied to the nominal weights assigned to each indicator. Intuitively, one may believe that the assignment of nominal weights to indicators would determine the influence each indicator has on the composite score. However, when indicators are correlated or have unequal variances, each indicator can

have a larger or smaller influence on the resulting composite score than suggested by the nominal weights (Stevens et al., 2000; Stevens & Aleamoni, 1986; Wang & Stanley, 1970).

The actual influence of each indicator on the overall composite is referred to as the "effective weight" of each indicator. If the effective weights are found to differ substantially from the nominal weights, this may suggest that the resulting composite score does not accurately reflect the intended goals for the accountability system, thus reducing the validity of the composite scores. It is critical to examine the alignment of nominal and effective weights to ensure the composite scores reflect the intended characteristics of schools and are interpreted appropriately.

Evaluating alignment between nominal and effective weights requires having a method to calculate effective weights. There is no unique way to define and thus operationalize the effective weight of an indicator because there are different ways to define how much "influence" each indicator has on the composite. Although there is no unique method to operationalize effective weights, we recommend using the following widely adopted method in educational measurement contexts (e.g., Baldwin, 2015; Schochet, 2008; Wang & Stanley, 1970). To formalize the definition and calculation of effective weights, let the composite indicator be C, where

$$C = \Sigma_{i=1}^{N} w_i Y_i. \quad (1)$$

Here there are N total indicators included in the composite, Y_i is the value of indicator i, and w_i is the nominal weight assigned to indicator i.

The total variance of the composite scores can be calculated based on the nominal weights, indicator variances, and indicator covariances as:

$$Var(C) = \sum_{i,j=1}^{N} w_i w_j Cov(Y_i, Y_j) = \sum_{i=1}^{N} w_i^2 Var(Y_i) + \sum_{i \neq j}^{N} w_i w_j Cov(Y_i, Y_j). \quad (2)$$

Here $Cov(\square)$ represents the covariance between two variables and $Var(\square)$ is the variance of a variable. The effective weight of each indicator is its contribution to the total variance of C, calculated as (Wang & Stanley, 1970):

$$d_i = w_i^2 Var(Y_i) + \sum_{j \neq i}^{N} w_i w_j Cov(Y_i, Y_j). \quad (3)$$

Examining Equation 3, there are three factors that will impact the effective weight of an indicator. All else equal, indicators with larger nominal weights, larger variances, or larger average covariances with other indicators will tend to have larger effective weights. In practice, these three factors combine, making it difficult to determine a priori which indicator will have the largest effective weight.

We can also consider two special cases. First, when the indicators are uncorrelated, the effective weight of each indicator reduces to:[2]

$$d_i = w_i^2 Var(Y_i). \quad (4)$$

Second, if indicators are standardized prior to creating the composite to reduce the influence of differential variances (e.g., Stevens & Aleamoni, 1986) the effective weight of each indicator reduces to:

$$d_i^z = w_i^2 + \sum_{j \neq i}^{N} w_i w_j Corr(Y_i^z, Y_j^z), \quad (5)$$

where Y_i^z is the standardized version of Y_i and $Corr(\square)$ is the Pearson correlation.

The sum of the d_i values will equal the total variance of C and the values will not in general be directly interpretable or comparable to the nominal weights w_i. To make the

[2] Equation 4 also highlights a discrepancy between the scale of nominal and effective weights. Assume the simplest case in which two uncorrelated and standardized indicators are combined so that each contributes uniquely to the variance of the composite. The effective weight of each variable will be w_i^2 rather than w_i. If the nominal and effective weights are rescaled each to sum to 100 (or another constant) the nominal and effective weights will only be equal if one of the nominal weights is 0 or if the nominal weights are identical. This discrepancy occurs because effective weights are computed relative to variances whereas nominal weights are applied to the original units of the variables. This is another reason that, in general, nominal and effective weights are not expected to be equal, although their relative magnitudes can be similar.

values more interpretable the effective weight can be converted to a percent metric using the equation

$$\delta_i = \frac{d_i}{\Sigma d_i} \times 100. \quad (6)$$

The δ_i will sum to 100 and represent the percent of total variance in C that is attributed to indicator Y_i. We refer to δ_i as the effective weight unless otherwise noted because in most cases the nominal weights are defined to represent a percentage for each indicator and will sum to 100% across indicators. If the nominal weights do not sum to 100 and instead sum to some arbitrary value T, the effective weights can be converted to a metric that is directly comparable to the nominal weights as

$$\delta_i^T = \delta_i \left(\frac{T}{100}\right). \quad (7)$$

Alternatively, the nominal weights could be converted to percentages that sum to 100 and compared directly to δ_i.

We prefer this operationalization of effective weights because it has been widely used in educational measurement contexts and has a straightforward interpretation, but we note three points to consider. First, this approach assumes the composite is a perfect linear combination of the indicators. In complex cases the composite may not be a linear combination of the indicators. If the composite is nearly a linear combination it may still be appropriate to use this approach. We illustrate such an example below based on the high-school SPF indicators. However, if the composite differs considerably from a linear combination this approach may not be appropriate.

Second, there are other ways effective weights could be operationalized and measured, which could lead to different conclusions. A simpler approach might use the correlation between each indicator and the composite to measure effective weights (see Domaleski (2019) for an overview and application to school accountability composites). However, this approach does not account for the correlation among indicators, which can

make indicators appear more influential than they are and does not produce values that can be directly compared to nominal weights. More complex approaches have also been proposed (e.g., Baldwin, 2015; Paruolo et al., 2013). Variable importance metrics in regression analysis (e.g., Grömping, 2015) provide a framework for conceptualizing the relative importance (and hence influence) for a set of predictor variables used to predict a criterion that could be adapted to define and operationalize effective weights. When the composite is a linear combination of the indicators, some variable importance metrics will be equivalent to the effective weights defined above. The Pratt index measure of variable importance in linear regression (Thomas et al., 1998), for example, will be equivalent to the effective weights described above when the criterion variable is an exact linear function of the predictor variables.

Third, the effective weight for an indicator could be negative if an indicator is negatively correlated with other indicators. While this is mathematically possible, we assume that all indicators being combined into a composite will be positively correlated because they are representing different facets of school quality, in which case all indicators will have positive effective weights.

The Colorado School Performance Framework (SPF)

The Colorado SPF produces an annual overall rating for each public school in the state. The SPF ratings are intended to provide information about how well each school is supporting students. The ratings are used to inform the development of customized supports for schools that are not supporting all students to succeed academically. The overall SPF rating is a composite score calculated based on academic achievement and academic growth indicators for elementary and middle schools and based on academic achievement, academic growth, and postsecondary and workforce readiness (PWR) indicators for high schools. Each of these indicators are themselves composite scores, which we discuss further below.

For elementary and middle schools, the achievement indicator is based on average scale scores from the Colorado Measures of Academic Success (CMAS) state standardized math and ELA tests administered in grades 3-8. Average scores for students in different grades, subjects, and subgroups are combined using a complex formula, resulting in a final overall achievement score for the school (Colorado Department of Education, 2020). The academic growth indicator is calculated using a similar process applied to student growth percentiles (SGP; Betebenner, 2009) based on CMAS performance for students in grades 4-8. Each indicator is included in the SPF calculation as a total percentage of points earned. As shown in Equation 8, these percentages are combined using nominal weights of 40% and 60% for achievement and growth, respectively, to arrive at an overall SPF composite score for each school:

$$SPF_{EM} = 40\left(\frac{Achievement\ pts.\ earned}{Achievement\ pts.\ eligible}\right) + 60\left(\frac{Growth\ pts.\ earned}{Growth\ pts.\ eligible}\right). \quad (8)$$

For high schools, the achievement indicator is based on average achievement on the PSAT for students in grades 9 and 10, and the growth indicator is based on median SGP from student PSAT/SAT scores in grades 9-11. The PWR indicator is based on average scores on the grade 11 SAT, dropout rates, matriculation rates, and graduation rates. The high school composite score is calculated as the weighted sum of achievement, growth, and PWR indicators, with weights of 30%, 40%, and 30%, respectively:

$$SPF_H = 30\left(\frac{Ach.\ pts.\ earned}{Ach.\ pts.\ eligible}\right) + 40\left(\frac{Growth\ pts.\ earned}{Growth\ pts.\ eligible}\right) + 30\left(\frac{PWR\ pts.\ earned}{PWR\ pts.\ eligible}\right). \quad (9)$$

As noted above, the achievement, growth, and PWR indicators are themselves composite scores based on additional subindicators. To illustrate how nominal and effective weights can be compared at the subindicator level we also investigate the PWR indicator as a composite. The PWR indicator is a composite score based on five subindicators: dropout

rates, graduation rates, SAT EBRW average scores, SAT Math average scores, and matriculation rates. The process used to construct the PWR composite from these five subindicators is complex and varies depending on the sample sizes of student subgroups within each school. For a school with sufficient student sample sizes to be eligible for all PWR points, the aggregation process is equivalent to using nominal weights of 9.2, 9.2, 4.6, 4.6, and 2.3, each divided by 30 (the total weight assigned to the PWR indicator), as follows:[3]

$$\begin{aligned} PWR\ Percent = & \frac{9.2}{30}\left(\frac{Dropout\ pts.earned}{Dropout\ pts.eligible}\right) + \frac{9.2}{30}\left(\frac{Grad.pts.earned}{Grad.pts.eligible}\right) \\ & + \frac{4.6}{30}\left(\frac{SAT\ EBRW\ pts.earned}{SAT\ EBRW\ pts.eligible}\right) \\ & + \frac{4.6}{30}\left(\frac{SAT\ math\ pts.earned}{SAT\ math\ pts.eligible}\right) \\ & + \frac{2.3}{30}\left(\frac{Matriculation\ pts.earned}{Matriculation\ pts.eligible}\right). \end{aligned} \quad (10)$$

These nominal weights sum to 30 (with a slight difference due to rounding) so that they can be compared directly to the nominal weights used for the overall high school SPF score above. The SPF rating system was designed so that, for example, the dropout subindicator can be interpreted as accounting for 9.2% of the overall high school SPF score.[4] If the nominal weights were scaled to sum to 100% for the PWR indicator alone, the nominal weight for the dropout subindicator would be (9.2/30)*100=30.67%. Additional details about the SPF calculations are provided in the Appendix.

The nominal weights and points assigned to each indicator in the SPF were determined by the State Board of Education based on recommendations from a panel of stakeholders consisting of district and other community organization personnel. The growth

[3] For schools with sample sizes below the minimum n-size, the number of eligible points and values of the nominal weights for the PWR composite can vary slightly. As a result, the PWR composite is not a perfect linear combination of the subindicators. However, for the illustrative purposes of this analysis, we assume that fixed nominal weights are applied to PWR subindicators for all schools and treat the PWR composite as if it were a linear combination. The actual PWR composite is nearly identical to the composite that would result from applying the nominal weights above, making this simplification reasonable.

[4] https://www.cde.state.co.us/accountability/pwr_factsheet

indicators were assigned larger nominal weights to communicate the importance of monitoring academic growth since growth is considered a better measure of how well schools are supporting student learning than average achievement, which is impacted by many out of school factors. The higher weight placed on growth is intended to communicate the importance of ensuring that all students are demonstrating growth and to construct a more valid measure of school quality for accountability purposes. The assumption of this claim is that the growth indicator will have a larger effective weight in the SPF composite scores. The weights for the PWR subindicators were selected to reflect the importance of different facets of PWR.

Data

We use publicly available SPF data for the 2022-23 academic year to calculate the effective weights and evaluate alignment between the nominal and effective weights of each indicator. The data provide the number of earned and eligible indicator and subindicator points, as well as the overall SPF composite for each school in Colorado. The data were retrieved from the CDE Schoolview site.[5] We computed the total percentages of points earned for each indicator or subindicator by dividing the total number of points earned by the total number of eligible points for each school. For consistency with the operational SPF scores, if a school enrolls students across multiple grade spans, for example both elementary grades (3-5) and middle grades (6-8) we consider the elementary grades to be one "school unit" and the middle grades to be a separate "school unit."

The analysis is restricted to schools with reported scores for all relevant indicators and subindicators. Schools with missing data for a given indicator or subindicator due to not meeting the minimum n-size requirements were excluded from the analysis. PWR subindicator scores were missing in some schools despite these schools having non-missing

[5] https://www.cde.state.co.us/schoolview/datafiles

PWR overall scores. The final analytic sample includes 1,651 elementary and middle school units (across 1,417 unique school buildings) and 420 high schools for the overall SPF analysis, and 414 high schools for the PWR sub-indicator analysis.

The calculation of operational SPF scores is based on additional business rules established by the state. For example, the state allows certain subgroups of students to earn a bonus point for reading proficiency. Other rules can be reviewed in Appendix A and Appendix B. For our illustrative purposes the analyses presented in this report rely on the reported eligible and earned points on the primary indicators (achievement, growth, and PWR) and subindicators (dropout, graduation, SAT performance, and matriculation). Results could differ slightly if applied to the operational scores.

Effective Weights of Indicators in the SPF

This section compares nominal and effective weights for the composite SPF scores among all schools and for the composite PWR scores among high schools. We first report results for elementary and middle schools and then for high schools.

Elementary/Middle School Indicators

Table 1 displays descriptive statistics for the percentage of points earned for each indicator across schools. The 1,651 school units in the sample earn on average 58% (SD=22.4%) of achievement points and 62.6% (SD=16.2%) of growth points, with a range of 25% to 100% of total possible points. The correlation between growth and achievement indicators is r=0.52, while the correlations of the indicators with the overall composite are r=0.86 and r=0.88 for achievement and growth, respectively. The variance of the achievement indicator (501.6) is nearly double the variance of the growth indicator (261.0), suggesting that the achievement indicator is likely to receive a higher effective weight relative to its nominal weight.

Table 1

Descriptive statistics of elementary and middle school achievement and growth indicators

Indicator	N Units	N Schools	Mean	SD	Variance	Min	Max
Achievement	1,651	1417	58.0	22.4	501.6	25	100
Growth	1,651	1417	62.6	16.2	261.0	25	100

Table 2 reports the nominal and effective weights, with effective weights scaled to sum to 100 for comparability to the nominal weights. Achievement has an effective weight of about 47% and growth has an effective weight of about 53%. The effective weight of achievement is larger than its nominal weight and the effective weight of growth is smaller than its nominal weight, but the relative magnitude of the nominal and effective weights for the indicators are consistent. As noted above, nominal and effective weights are not expected to be identical. However, the effective weight of growth is only about 10% larger than the effective weight of achievement, while the nominal weight of growth was intended to be 50% larger than the nominal weight of achievement.

Table 2

Nominal and effective weights of elementary and middle school indicators

Indicator	Nominal Weight	Effective Weight
Achievement	40.0%	47.4%
Growth	60.0%	52.6%

High School Indicators

Table 3 displays descriptive statistics for the high school indicators. Among the 420 schools with complete data, the percentage of points earned by indicator ranges from 25% to 100% for achievement and PWR, and between 25% and 91.7% for growth. On average, schools earn the highest percentage of PWR points, followed by growth and then achievement. The variances of the three indicators differ substantially. Achievement has the highest variance, followed by PWR and then growth. The variance of achievement is nearly

three times larger than the variance of growth. Table 4 displays the correlations among the three indicators and the overall weighted composite across high schools. Achievement has the highest correlation with the composite, although all three indicators are highly correlated with the composite. The correlations among indicators are similar, with a correlation of r=0.69 between achievement and each other indicator and r=0.60 between growth and PWR.

Table 3

Descriptive statistics of high school achievement, growth, and PWR indicators

Indicator	N Schools	Mean	SD	Variance	Min.	Max
Achievement	420	52.1	21.8	475.4	25	100
Growth	420	57.3	12.9	167.5	25	91.7
PWR	420	66.8	19.2	368.1	25	100

Table 4

Correlations between high school SPF indicators

	Composite	Achievement	Growth	PWR
Composite	1.00			
Achievement	0.91	1.00		
Growth	0.85	0.69	1.00	
PWR	0.87	0.69	0.60	1.00

Table 5 displays the nominal weights and effective weights for the three high school indicators. The weights in each column sum to 100%. The effective weights are approximately 39% for achievement, 29% for growth, and 33% for PWR. The effective weights for achievement and PWR are larger than their nominal weights, while the effective weight of growth is smaller than its nominal weight. In contrast to the elementary and middle school indicators, the high school effective weights do not maintain the same relative ordering as the nominal weights: Growth has the largest nominal weight, but it has the smallest effective weight of the three indicators.

Table 5

Nominal and effective weights of high school indicators

Indicator	Nominal Weight	Effective Weight
Achievement	30.0%	38.8%
Growth	40.0%	28.7%
PWR	30.0%	32.5%

PWR Subindicators

Table 6 displays descriptive statistics for the PWR subindicators across the 414 high schools with complete PWR data. Schools earned between 25% and 100% of eligible points across each of the five subindicators. Schools earned the greatest percentage of graduation rate points and the lowest percentage of matriculation points, on average. Most relevant for the comparison of nominal and effective weights, the variances of the subindicators were more similar than in the examples above. The graduation subindicator had the largest variance while the SAT Math subindicator had the smallest variance. Table 7 presents correlations among the subindicators and the composite PWR indicator. All subindicators were strongly positively correlated with the composite and moderately to strongly positively correlated with other subindicators.

Table 6

Descriptive statistics of high school PWR sub-indicators

Subindicator	N Schools	Mean	SD	Variance	Min.	Max
Dropout	414	68.43	24.18	584.68	25	100
Graduation	414	78.52	24.61	605.73	25	100
SAT EBRW	414	52.77	23.14	535.42	25	100
SAT Math	414	48.13	21.91	479.95	25	100
Matriculation	414	48.55	23.55	554.80	25	100

Table 7

Correlations between sub-indicators on high school PWR

	PWR Total	Dropout	Grad	EBRW	Math	Matriculation
PWR Total	1.00					
Dropout	0.92	1.00				
Graduation	0.86	0.71	1.00			
SAT EBRW	0.76	0.63	0.47	1.00		
SAT Math	0.79	0.63	0.52	0.89	1.00	
Matriculation	0.67	0.51	0.53	0.47	0.55	1.00

Note. The PWR Total variable is the official PWR composite for each school, not a linear weighted composite using the nominal weights (as explained above). Correlation between a nominally weighted PWR composite and the official PWR composite is r=0.998. Using either version results in the same substantive conclusions.

Table 8 displays the nominal and effective weights of each subindicator. All weights are scaled to sum to 30 for comparability and apply only to the PWR composite (not the overall SPF composite). Because the PWR composite gets re-weighted as part of the overall SPF composite, we cannot say that the dropout subindicator has an effective weight of 10.1% of the overall SPF composite. However, we can compare this effective weight directly to the nominal weight of 9.2% that is also scaled to sum to 30. The effective weights of the subindicators were close to the nominal weights. Dropout and graduation subindicators had slightly larger effective weights relative to their nominal weights while the remaining subindicators had smaller effective than nominal weights.

Table 8

Nominal and effective weights of PWR sub-indicators

Subindicator	Nominal Weight	Effective Weight
Dropout	9.2%	10.1%
Graduation	9.2%	9.8%
EBRW	4.6%	4.2%
Math	4.6%	4.1%
Matriculation	2.3%	1.8%

Note. These effective weights are calculated assuming the nominal weights would be used directly to create the PWR composite.

Standardized Indicators

For an additional reference point, we re-calculated all effective weights after converting the indicators and subindicators to standardized z-scores.[6] Standardizing indicators before combining them into a weighted composite maintains the correlation among indicators but reduces the influence of unequal variances across indicators because all indicators have equal variance after standardizing. The effective weights calculated based on standardized indicators are reported in Table B1 and were significantly closer to the nominal weights than the effective weights calculated using the original indicators reported above. This finding suggests the observed discrepancy between nominal and effective weights above is caused by unequal variances across indicators and supports the claim that standardizing indicators prior to forming the composite is one way to increase the alignment between nominal and effective weights. However, there are downsides to standardizing indicators. Standardized indicators may be more difficult to explain in some contexts, because standardized indicators can be negative. Standardizing also reduces comparability across time, because standardized indicators depend on the distribution of scores across schools and hence the equation for standardizing (and thus computing overall accountability scores) can vary from year to year.

Discussion

This investigation of nominal and effective weights in SPF scores demonstrated that the effective weights of indicators, defined as the proportion of composite variance explained by each indicator, do not always align with the assigned nominal weights. In the elementary and middle school SPF, the effective weight of growth was larger than that of achievement, but not by as much as stakeholders intended based on the assigned nominal weights. In the

[6] A z-score is a measurement of the number of standard deviations a data point is from the mean. A positive z-score indicates a data point above the mean; a negative z-score indicates a data point below the mean.

case of high school SPF indicators, the relative magnitude of the effective weights was counter to what the nominal weights suggest – growth was assigned the largest nominal weight of the three indicators, but it had the smallest effective weight. The nominal and effective weights for the PWR subindicators were well aligned. In general, the indicators and subindicators with more variance tended to receive larger effective weights than suggested by their nominal weights, while those with less variance tended to receive smaller effective weights than suggested by their nominal weights.

There are some limitations to the scope of our investigation to note. First, we did not compare results based on alternative methods for defining and calculating effective weights. Alternative ways of defining and calculating effective weights may be more appropriate when the composite score is not a direct linear combination of the indicators or when a different operationalization of "effective weight" is warranted. Second, we did not consider methods that could be used to determine the nominal weights necessary to achieve a desired set of effective weights. Although possible, this makes the process of setting nominal weights statistically more complicated and may not be desirable from a policy perspective that values transparency and explainability. Third, we demonstrated these results using a single state system (the Colorado SPF system). However, the concepts and approach to comparing nominal and effective weights we use could be applied to any state accountability (or other) system relying on weighted composite scores.

We should emphasize that the differences observed between nominal and effective weights do not render SPF composite scores incorrect or invalid for their intended purposes. Unlike some contexts in which there is a latent construct that a manifest score or indicator is intended to represent, there is no underlying "true" SPF score the composite is intended to represent. If all SPF data and calculations are implemented properly, then by definition the resulting SPF score is correct. An evaluation of the validity of SPF ratings would require

considering how well the SPF ratings support their intended uses and interpretations, which cannot be determined solely by comparing the nominal and effective weights. Effective weights provide only one lens through which to understand how indicators influence a composite score.

Nonetheless, those designing and implementing the SPF system should be aware of the differences between nominal and effective weights and take these differences into account when evaluating or revising the system. If people using SPF scores rely on the nominal weights to understand the meaning of the composite scores and assume the nominal weights signify how much influence each indicator has on a school's rating, this may lead to inaccurate inferences. The nominal weights used to construct SPF composites may need to be revised if Colorado intends for growth to have a larger influence on the overall variability of SPF scores than other indicators, in terms of variance. If the goal is to achieve a specific set of effective weights, nominal weights could be selected that lead to effective weights equal to the desired relative magnitudes. Alternatively, if the goal is to increase alignment between nominal and effective weights, indicators could be transformed or constructed differently to equalize variances prior to weighting them. A full consideration of alternative approaches to setting nominal weights is beyond the scope of this report.

References

Baldwin, P. (2015). Weighting components of a composite score using naïve expert judgments about their relative importance. *Applied Psychological Measurement*, *39*(7), 539–550. https://doi.org/10.1177/0146621615584703

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

Colorado Department of Education. (2020). *2019 District and School Performance Framework Calculation Guidebook.* https://www.cde.state.co.us/accountability/district-and-school-performance-framework-calculation-guidebook-2019

Domaleski, C. (2019, May 24). Understanding nominal and effective weights in school accountability systems. *CenterLine Video Blog*. https://www.nciea.org/blog/understanding-nominal-and-effective-weights-in-school-accountability-systems/

Erwin, B., Francies, C., Pechota, D., & McCann, M. (2021). *50-state comparison: State's school accountability systems*. Education Commission of the States. https://www.ecs.org/50-state-comparison-states-school-accountability-systems/

Grömping, U. (2015). Variable importance in regression models. *WIREs Computational Statistics*, *7*(2), 137–152. https://doi.org/10.1002/wics.1346

Paruolo, P., Saisana, M., & Saltelli, A. (2013). Ratings and rankings: Voodoo or science? *Journal of the Royal Statistical Society Series A: Statistics in Society*, *176*(3), 609–634. https://doi.org/10.1111/j.1467-985X.2012.01059.x

Schochet, P. Z. (2008). *Technical methods report: Guidelines for multiple testing in impact evaluations* (NCEE 2008-4018). National Center for Education Evaluation and

Regional Assistance: Institute of Education Sciences. https://ies.ed.gov/ncee/pdf/20084018.pdf

Stevens, J., & Aleamoni, L. M. (1986). The role of weighting in the use of aggregate scores. *Educational and Psychological Measurement*, *46*(3), 523–531. https://doi.org/10.1177/0013164486463003

Stevens, J., Estrada, S., & Parkes, J. (2000, April 24). *Measurement issues in the design of state accountability systems*. Annual Meeting of the American Educational Research Association, New Orleans, LA. https://eric.ed.gov/?id=ED445026

Thomas, D. R., Hughes, E., & Zumbo, B. D. (1998). On variable importance in linear regression. *Social Indicators Research*, *45*, 253–275. https://doi.org/10.1023/A:1006954016433

Wang, M. W., & Stanley, J. C. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research*, *40*(5), 663–705. https://doi.org/10.3102/00346543040005663

Appendix A: SPF Scoring Details

SPF Scoring Guide

Draft Scoring Guide for 2024 District/School Performance Frameworks

Performance Indicator	Measure/Metric	Rating	Point Value		
Academic Achievement & ELP On Track Growth	*The district or school's mean scale score (or percent On Track) was*: see tables below for actual values*		*All Students*	*Each Disaggregated Group*	*ELP On Track Growth*
	• at or above the 85th percentile	Exceeds	8	1.00	2.0
	• at or above the 50th percentile but below the 85th percentile	Meets	6	0.75	1.5
	• at or above the 15th percentile but below the 50th percentile	Approaching	4	0.50	1.0
	• below the 15th percentile	Does Not Meet	2	0.25	0.5
	Students Previously Identified for a READ Plan (bonus point)				
	• CMAS ELA Mean scale score at or above 725 (Approaching Expectations cut-score)		1 bonus point		
Academic Growth	*Median Growth Percentile was:*		*All Students*	*Each Disaggregated Group*	*ELP*
	• at or above 65	Exceeds	8	1.00	2.0
	• at or above 50 but below 65	Meets	6	0.75	1.5
	• at or above 35 but below 50	Approaching	4	0.50	1.0
	• below 35	Does Not Meet	2	0.25	0.5
Postsecondary and Workforce Readiness	*Mean CO SAT Evidence-Based Reading and Writing (EBRW) scale score was**:*		*All Students*	*Each Disaggregated Group*	
	• at or above 554.7	Exceeds	4	1.00	
	• at or above 501.3 but below 554.7	Meets	3	0.75	
	• at or above 458.0 but below 501.3	Approaching	2	0.50	
	• below 458.0	Does Not Meet	1	0.25	
	*Mean CO SAT Math scale score was**:*		*All Students*	*Each Disaggregated Group*	
	• at or above 544.6	Exceeds	4	1.00	
	• at or above 488.0 but below 544.6	Meets	3	0.75	
	• at or above 439.9 but below 488.0	Approaching	2	0.50	
	• below 439.9	Does Not Meet	1	0.25	
	Dropout Rate: The district or school dropout rate was (of all schools in 2017):		*All Students*	*Each Disaggregated Group*	
	• at or below 0.5%	Exceeds	8	2.0	
	• at or below 2.0% but above 0.5%	Meets	6	1.5	
	• at or below 5.0% but above 2.0%	Approaching	4	1.0	
	• above 5.0%	Does Not Meet	2	0.5	
	Matriculation Rate (of all schools in 2018):		*All Students*		
	• at or above the 75.8%	Exceeds	4		
	• at or above 61.1% but below 75.8%	Meets	3		
	• at or above 46.8% but below 61.1%	Approaching	2		
	• below 46.8%	Does Not Meet	1		
	Graduation Rate and Disaggregated Graduation Rate (Best of 4-, 5-, 6-, or 7-year):		*All Students*	*Each Disaggregated Group*	
	• at or above 95.0%	Exceeds	8	2.0	
	• at or above 85.0% but below 95.0%	Meets	6	1.5	
	• at or above 75.0% but below 85.0%	Approaching	4	1.0	
	• below 75.0%	Does Not Meet	2	0.5	

Academic Achievement: Mean Scale Score by Percentile Cut-Points

The Academic Achievement indicator reflects achievement as measured by the mean scale score on Colorado's standardized assessments. The presented targets for the Achievement indicators have been established utilizing baseline year data.*

	English Language Arts & EBRW for CO PSAT			Mathematics			Science (Draft)		
Percentile	Elementary	Middle	CO PSAT	Elem	Middle	CO PSAT	Elem	Middle	High
15th percentile	722.3	724.1	423.5	719.1	716.5	413.0	713.4	714.6	720.2
50th percentile	739.5	740.1	461.1	734.3	731.2	448.4	733.1	732.5	734.4
85th percentile	755.9	757.3	505.0	751.9	746.2	491.0	750.2	748.8	745.4

Percent of Students On Track for ELP Growth Targets

	ELP On Track Growth		
Percentile	Elem	Middle	High
15th percentile	51.7%	8.9%	12.1%
50th percentile	64.1%	18.0%	21.1%
85th percentile	76.3%	31.5%	29.8%

Total Possible Points by Performance Indicator

Indicator	Total Possible Points	Elementary/Middle	High/District
Achievement	36 points (8 per subject for all students, 4 per subject by disaggregated group)	40%	30%
Growth	28 total points (8 per subject for all students, 4 per subject by disaggregated group, 2 for ELP growth, 2 for ELP On Track Growth)	60%	40%
Postsecondary Readiness	52 total points (16 for graduation, 4 for matriculation, 16 for dropout, 8 per CO SAT subject)	not applicable	30%

Cut-Points for Each Performance Indicator

	Cut-Point: The district or school earned...of the points eligible.	
Achievement; Growth; Postsecondary Readiness	• at or above 87.5%	Exceeds
	• at or above 62.5% but below 87.5%	Meets
	• at or above 37.5% but below 62.5%	Approaching
	• below 37.5%	Does Not Meet

Cut-Points for Plan/Category Type Assignment

	District	*School*	*Accreditation Category/Plan Type*
Total Framework Points	74.0%	not applicable	Accredited w/Distinction (District only)
	56.0%	53.0%	Accredited (District) or Performance Plan (School)
	44.0%	42.0%	Accredited w/Improvement Plan (District) or Improvement Plan (School)
	34.0%	34.0%	Accredited w/Priority Improvement Plan (District) or Priority Improvement (School)
	25.0%	25.0%	Accredited w/Turnaround Plan(District) or Turnaround Plan (School)

* School data used as baseline: 2016 for CMAS & CoAlt ELA & Math (g3-8). 2019 for CO PSAT & CoAlt EBRW/ELA & Math (g9-10). 2023 for ELP On Track to Proficiency Growth as planned prior to the COVID-19 pandemic.

** 2019 school data used as baseline for CO SAT & CoAlt EBRW/ELA & Math (g11).

November 16, 2023

Source: https://www.cde.state.co.us/accountability/performanceframeworks

Table A1

SPF calculations for elementary and middle schools

Achievement 40%		Growth 60%	
All students	8pts	All students	8pts
Each disaggregated group	1pt	Each disaggregated group	1pt
ELP on track growth	2pts	ELP	2pts
Students previously identified for READ plan	1 BONUS point		
36 points (8 per subject for all students, 4 per subject by disaggregated group)		28 total points (8 per subject for all students, 4 per subject by disaggregated group, 2 for ELP growth, 2 for ELP on track growth)	

Table A2

SPF calculations for high schools

Achievement 30%		Growth 40%		PWR 30%		
All students	8pts	All students	8pts	Dropout - 9.2%		
Each disaggregated group	1pt	Each disaggregated group	1pt		All students	8pts
ELP on track growth	2pts	ELP	2pts		Each disaggregated group	2pts
Students previously identified for READ plan	1 BONUS point			Graduation - 9.2%		
					All students	8pts
					Each disaggregated group	2pts
				SAT - EBRW - 4.6%		4.60%
					All students	4pts
					Each disaggregated group	1pt
				SAT - Math - 4.6%		4.60%
					All students	4pts
					Each disaggregated group	1pt
				Matriculation - 2.3%		2.30%
					All students	4pts
36 points (8 per subject for all students, 4 per subject by disaggregated group)		28 total points (8 per subject for all students, 4 per subject by disaggregated group, 2 for ELP growth, 2 for ELP on track growth)		52 total points (16 for dropout, 16 for graduation, 8 per CO SAT subject, 4 for matriculation)		

Table A3

Calculations for the high school PWR indicator

Achievement 30%		Growth 40%		PWR 30%		
All students	8pts	All students	8pts	Dropout - 9.2%		
Each disaggregated group	1pt	Each disaggregated group	1pt		All students	8pts
ELP on track growth	2pts	ELP	2pts		Each disaggregated group	2pts
Students previously identified for READ plan	1 BONUS point			Graduation - 9.2%		
					All students	8pts
					Each disaggregated group	2pts
				SAT - EBRW - 4.6%		4.60%
					All students	4pts
					Each disaggregated group	1pt
				SAT - Math - 4.6%		4.60%
					All students	4pts
					Each disaggregated group	1pt
				Matriculation - 2.3%		2.30%
					All students	4pts
36 points (8 per subject for all students, 4 per subject by disaggregated group)		28 total points (8 per subject for all students, 4 per subject by disaggregated group, 2 for ELP growth, 2 for ELP on track growth)		52 total points (16 for dropout, 16 for graduation, 8 per CO SAT subject, 4 for matriculation)		

Appendix B: Standardized Indicator Effective Weights

Table B1

Nominal and effective weights for original and standardized indicators

Level	Indicator	Nominal Weight	Original Effective Weight	Standardized Effective Weight
E/M				
	Achievement	40.0%	47.4%	37.0%
	Growth	60.0%	52.6%	63.0%
High School				
	Achievement	30.0%	38.8%	30.3%
	Growth	40.0%	28.7%	40.7%
	PWR	30.0%	32.5%	29.0%
PWR				
	Dropout	9.2%	10.1%	10.0%
	Graduation	9.2%	9.8%	9.4%
	EBRW	4.6%	4.2%	4.3%
	Math	4.6%	4.1%	4.5%
	Matriculation	2.3%	1.8%	1.8%