Revisiting N Size

Evaluating Outcomes on the School and District Performance Frameworks Relative to N Size

Elena Diaz-Bilello, Center for Assessment

Executive Summary

This report brief revisits concerns articulated by stakeholders that the Colorado Growth Model (CGM) and the overall ratings on the performance frameworks appear to be "biased" against large districts and schools, and that small institutions appear to be over-represented at the tails of the growth and final ratings distributions.

Although the set of analyses reviewing outcomes on the performance frameworks from 2008-2009 to 2011-2012 presented in this report appears to support stakeholder concerns, the information presented highlights a different issue: smaller institutions in general display higher levels of variability in performance regardless of whether the metric reflects outcomes from the CGM or other status based measures such as proficiency rates and ACT scores. Additionally, higher levels of variability found in smaller institutions relative to larger institutions would also be detected regardless of the growth model employed (e.g., using a value-added model) and regardless of whether outcomes for smaller institutions are being reviewed in the education or an entirely different sector. In education systems, this large degree of performance variability found in smaller institutions may be attributed to many factors, including:

- Smaller schools with mission specific goals (e.g., serving dropout students or gifted and talented students) are more likely to attract and recruit students sharing similar academic profiles at the high or the low end of the performance range;

- Population shifts are likely to have a higher performance impact on smaller institutions; and,

- Large gains or declines can be triggered by the performance shifts of a few students at smaller institutions.

Considering that more information from students is available to better evaluate the performance taking place in larger institutions and these institutions are less likely than smaller institutions to exhibit large performance shifts over time, recommendations are forwarded in this paper to the Colorado Department Education (CDE) to ensure that all of the evidence gathered in the framework sufficiently allow for well supported inferences to be made about academic performance taking place in smaller institutions. These recommendations include: ensuring that only pooled data are used; reviewing the stability of performance each year to determine whether smaller institutions require an automatic reconsideration of ratings achieved; and soliciting additional academic evidence of learning progress where only status metrics are available or no metrics are available to evaluate the performance of an institution. Further, this paper encourages CDE to consider additional opportunities for larger districts to achieve distinction by drawing on other indicators outside of the framework that are aligned with the state's goal of preparing all students for post-secondary and workforce readiness.

Introduction

This report brief represents the first in a three-part series examining three issues related to outcomes from the school and district performance frameworks. These issues include:

1. The relationship between a school and district's size and MGPs, growth ratings and overall ratings achieved on the frameworks each year;

2. The relationship between poverty, ELL, minority status and a school or district's ratings by key indicator and their overall ratings; and,

3. The stability of ratings achieved by schools and districts over time (2009-2012) on status, growth and overall ratings on the framework.

This first brief addresses the stakeholder concern that the median growth percentiles generated under the Colorado Growth Model (CGM) and the final performance ratings assigned to schools over-identify or single out smaller schools at the tails of the performance distribution, and that the CGM and the design of the overall frameworks are less likely to recognize the performance gains made by larger institutions. In this paper, these concerns are revisited from the perspective that the observed pattern found between n size and outcomes typically occurs for both proficiency based and CGM based metrics, and also occurs regardless of sector or field examined. A set of recommendations are presented in the discussion as considerations for ensuring that performance judgments – particularly for small schools and districts – are made using the best set of available information and evidence to inform those judgments.

Data Reviewed

District and school n size was examined by level (elementary, middle and high) relative to the following areas:

- Median growth percentiles and accompanying growth ratings achieved in each subject area on the one year and three year frameworks in 2009-2012;

- Percent proficient and above performance achieved in each subject area on the one year and the three year frameworks in 2012;

- ACT and "best" graduation rate performance achieved by high schools in 2012; and,

- Overall ratings achieved on the frameworks in 2009-2012 prior to adjusting ratings based on reconsideration requests.

Since the examination of outcomes on the achievement and post-secondary metrics on the frameworks was conducted for comparative and illustrative purposes, these outcomes were limited to the most current 2012 school year data set. The n sizes reflected with the median growth percentiles represent the total number of test takers with growth scores in each subject area. The n sizes reflected relative to the final ratings on the frameworks represent the number of test takers counted under math achievement. The n size for math serves as a proxy for the

entire population of test takers at a given school who are largely driving all outcomes on the frameworks. Math was selected over reading and writing since this subject typically reports larger numbers of test takers relative to the language arts subjects. The proxy n size for the test taking population was divided into four groups each year for the school performance frameworks (SPFs) as follows: 1) up to 182 students, 2) 183 – 395 students; 3) 396 – 603 students; and 4) more than 604 students. The test taking population was segmented as follows for the district performance frameworks (DPFs): 1) up to 350 students, 2) 351 – 3,000 students, 3) 3001-7,000 students, and 4) more than 7001 students.

Summary of Findings

Relationship of n size with MGPs and growth ratings by level, subject and year for SPFs

The scatter plots in Attachment 1 present the relationship between the n size of students with reading and math growth scores relative to MGP outcomes and growth ratings received. The plots were generated by level (elementary, middle and high), subject area and for each year. Each scatter plot generally conforms to a funnel shaped distribution where schools at the larger end of the size distribution (to the right of each plot) are more likely to approximate performance at the state level. More specifically:

- For schools located well below the reference line representing the average n size of students with growth scores, the scatter plots suggest a larger degree of variability in median percentiles achieved relative to those schools and districts above the reference line.

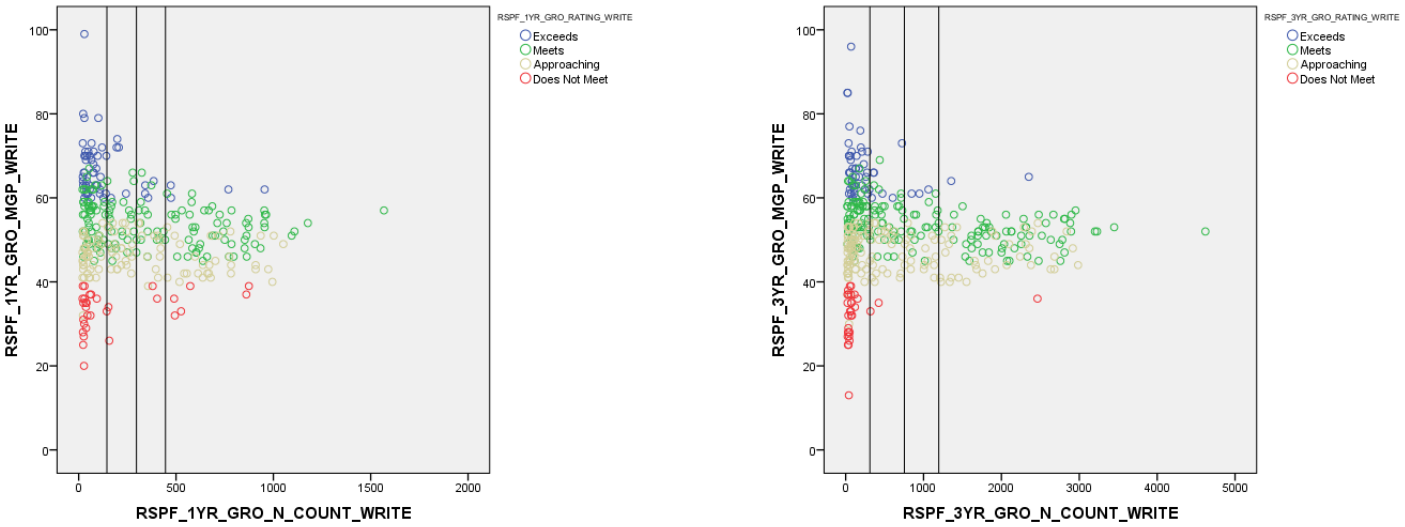
- For the vast majority of schools located just below and above the mean, growth performance is still variable, but with more schools just above the mean exhibiting growth within the 40th to the 60th percentile range and subsequently rated as "approaching" and "meets".

- For the small group of schools with an n size located well above the mean, these large schools tend to exhibit performance that range predominantly between the 40th and the 60th percentile and subsequently rated as "approaching" and "meets".

Overall, the set of scatter plots in Attachment 1 indicate that in each year reviewed, the MGPs tend to gravitate toward the state median as a function of n size, and that growth performance tends to be more variable at the lower end of the n distribution. These patterns are more pronounced in the middle and high schools where the n size distributions are considerably larger than the elementary schools. However, these funnel patterns exhibited by the scatter plots are not unique to the use of the CGM, but also occur using other growth modeling approaches and using other metrics driving the frameworks.

To illustrate, results from two sets of outcomes in 2012 for writing are compared. Figure 1 presents growth achieved by all high schools in 2012 in writing based on the one and three year framework results.

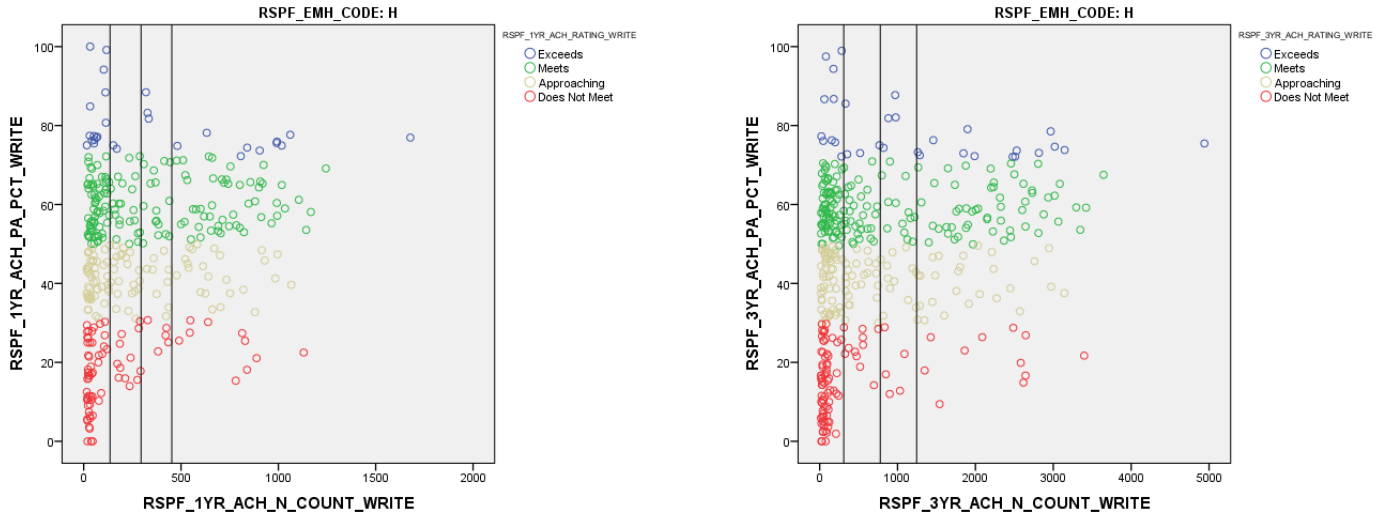
Figure 1. High schools growth by n size on 1 and 3 year frameworks



In each plot under Figure 1, the x-axis represents the n size of students with writing growth scores by schools and the y-axis presents the MGPs earned by schools on the writing test. The legend in each figure represents the growth ratings received on the writing test. Schools located below the first reference line in each figure represent schools with n sizes falling below .5 standard deviation from the mean, the second reference line represents the mean, and schools located well above the third reference line represent schools with an n size located above .5 standard deviation from the mean.

Figure 2 presents the proficiency performance achieved by all high schools on writing in 2012 on the one and the three year frameworks. In these figures, the first reference line represents schools with n sizes below .5 standard deviation from the mean, the second reference line represents the mean, and the third reference line represents the point located .5 standard deviation above the mean. The legends in Figures 3 and 4 reflect the growth ratings earned by high schools on achievement for writing.

Figure 2. High schools achievement by n size on 1 and 3 year frameworks



Similar to the patterns found for the MGPs, smaller schools located below the mean n size exhibit more variability in proficiency achieved relative to schools with considerably larger populations (above the third reference line). The larger entities with proficiency rates below 30 percent consist largely of alternative high schools. Since ACT scores and graduation rates are also evaluated for high schools, outcomes from these metrics were also reviewed in the 2012 year for high schools. Figures 4 and 5 present the results for high schools on the 1 year and 3 year frameworks on the ACTs and "best" graduation rate outcomes. In each plot, the first reference line represents the point where the n size falls .5 standard deviation below the mean, the second line represents the mean, and the third line represents the n size point located .5 standard deviation above the mean.

Figure 3. High schools achievement on ACTs

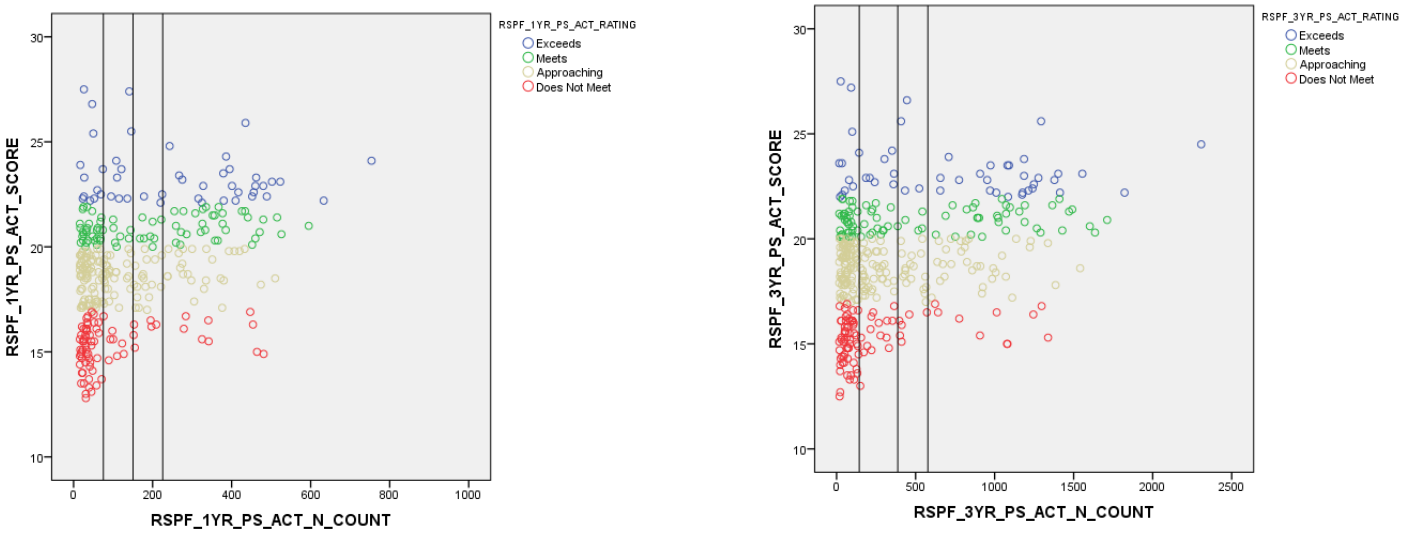
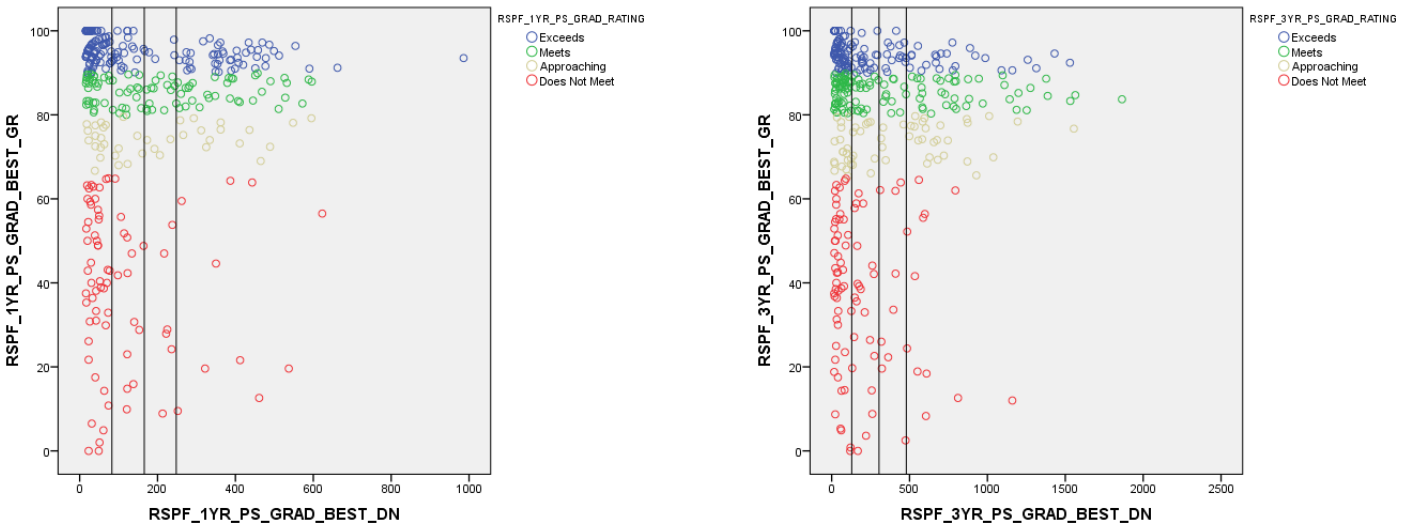


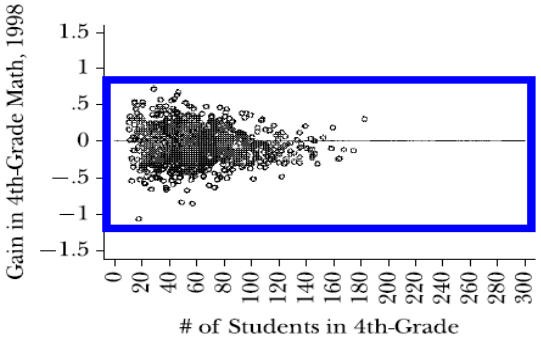
Figure 4. High schools and "best" graduation rate



Again, similar to the funnel pattern exhibited by the MGP and proficiency plots, the smaller schools in general tend to exhibit a wider range of performance on the ACTs and graduation rates achieved, and also appear at higher numbers at the tails of each distribution. However, compared to other status based metrics, the "best" graduation rate distribution is more highly skewed by the presence of alternative high schools that are largely located at the lower end of the scale.

Not only do these small and large n size performance patterns hold across the different metrics employed within the framework, but they also hold across different types of growth modeling approaches used in the field. Figure 5 illustrates findings from a study conducted by Kane and Staiger (2002) examining variability in performance gains found using a value-added model (VAM) relative to the n size of fourth grade students taking the math test across schools in North Carolina. As indicated by Figure 5, performance gains tend to gravitate toward the center of the distribution as the n size increases along the x-axis.

Figure 5. VAM results and n size



Source: Kane & Staiger, 2002

This funnel pattern found between n size and outcomes would still persist if other growth modeling approaches are considered for the framework since other regression based growth modeling approaches such as VAMs are known to be highly correlated with the CGM (see Briggs & Betebenner, 2009; Wright, 2010). Taken together, the plots in Figures 2 through 5 suggest that the CGM in-itself is not systematically biased against large schools, but rather that smaller schools, due to many factors, are more likely to exhibit more variable patterns of performance relative to larger schools on growth and other status based metrics. Factors contributing to the large variability in performance for smaller schools may include but are not limited to the following:

- Smaller schools with mission specific goals[1] (e.g., serving dropout students or gifted and talented students) are more likely to attract and recruit students sharing similar academic profiles at the high or the low end of the performance range;

- Population shifts are likely to have a higher performance impact on small relative to large schools; and,

- Large gains or declines can be triggered by the performance movements of a few students at small schools relative to large schools.

As noted by Wainer (2009), a finding that smaller schools are generally overrepresented at the tails of a distribution is "exactly as expected, since smaller schools will show greater variation in performance and empirically will show up where we look" (pg. 13). In the next section, the n size relationship to growth outcomes is also reviewed for the DPFs.

Relationship of n size with MGPs and growth ratings by level, subject and year for DPFs

Similar to the results found on the SPF scatter plots, the DPF scatter plots in Attachment 2 suggest:

- Higher levels of growth performance variability appear to be associated with smaller n sizes by subject area and level at the district level; and,

- Since the DPFs are associated with a considerably fewer number of units at each level (elementary, middle, and high), the vast majority of districts are located below the mean n size designated by the vertical line in each graph.

Compared to the SPF scatter plots in Attachment 1, the funnel pattern is more pronounced for the DPFs due to the fact that the vast majority of districts are located below the mean n size. The handful of school districts evaluated with much larger n sizes generally exhibit MGPs located above the 40th and below the 60th percentile.

[1] In the case of the few larger schools that recruit students based on a specific mission, these entities typically represent alternative education schools (e.g., R-5 in Mesa County and Emily Griffith Technical School in Denver) and exhibit lower levels of performance than schools of similar size as assessed by the regular school performance frameworks. However, these schools are evaluated separately using the Alternative Education Campus Performance Framework and are typically not compared against traditional high schools.

To illustrate once again that this small n size phenomenon would also be observed on the proficiency based metrics on the district frameworks, Figure 6 presents the 2012 MGPs achieved at the elementary level across all school districts in math, and Figure 7 presents the 2012 percent proficient and above achieved in the same subject area. The first reference line in each graph represents the average n size for the entire population. The second reference line represents all districts located above .5 standard deviation from the mean n size.

Figure 6. Elementary level district growth by n size on one and three year frameworks

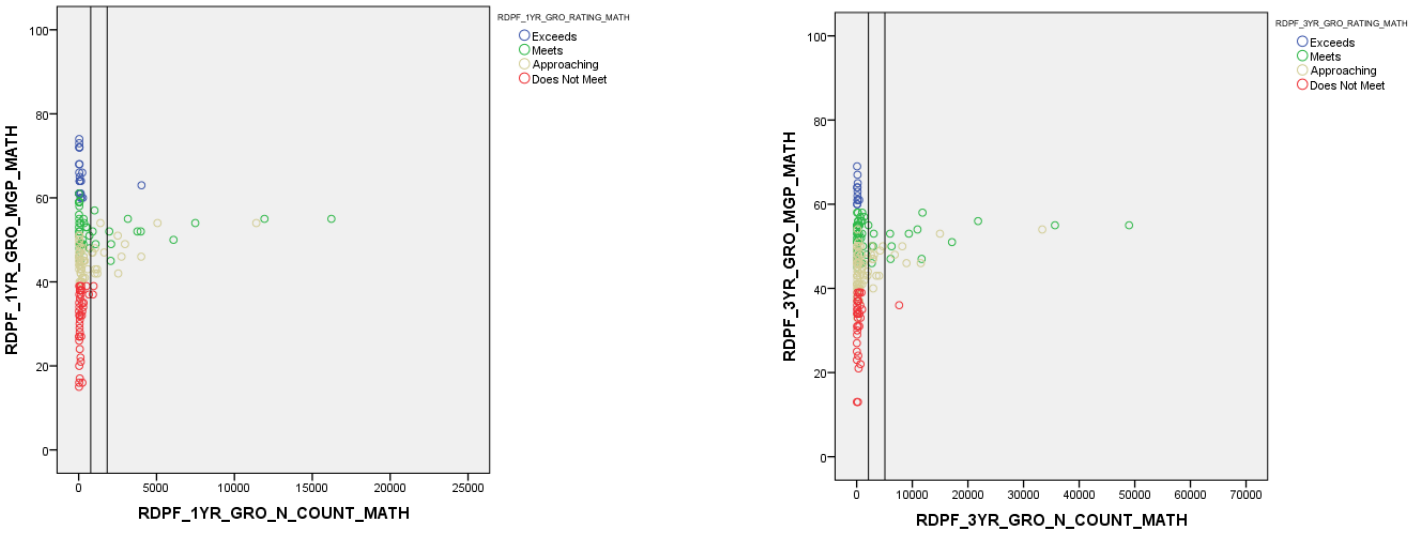
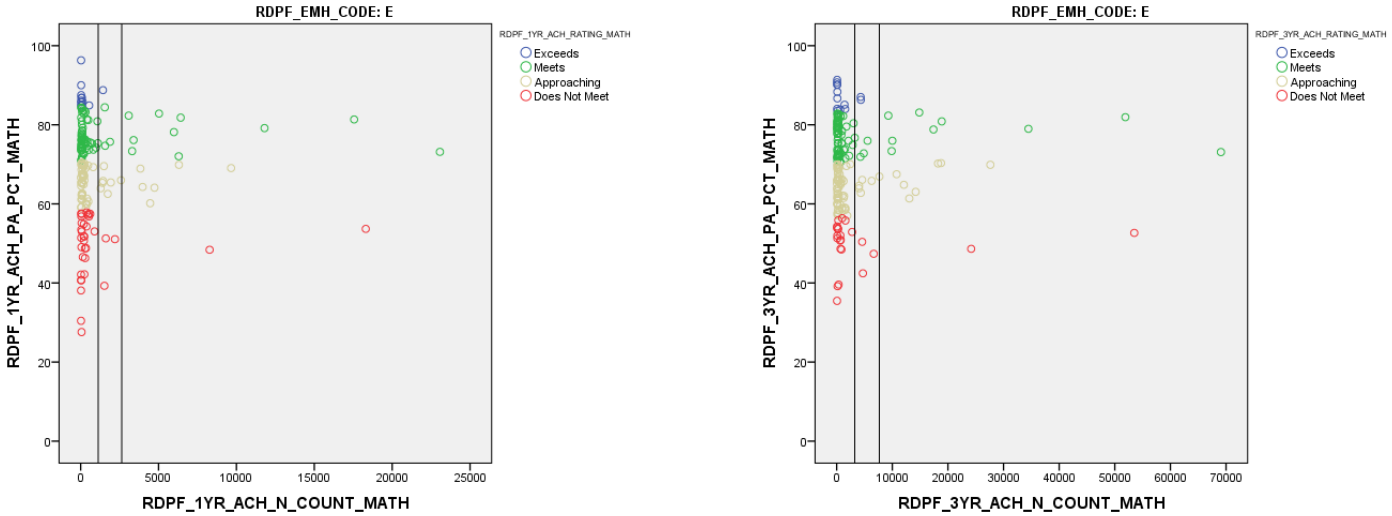


Figure 7. Elementary level district achievement by n size on one and three year frameworks



As highlighted by each plot, the vast majority of districts located below the mean n size appear at the tails of the performance distribution at higher proportions than those districts located above.5 standard deviation above the mean n size on either growth or proficiency. The ACT and graduation performance for high schools at the DPF level are presented in Figures 8 and 9 and also exhibit similar patterns found for growth and proficiency. For each scatter plot presented

in Figures 8 and 9, the first reference line represents the mean of the population and second reference line represents the point located .5 standard deviation above the mean.

Figure 8. ACT performance at the district level on one and three year frameworks

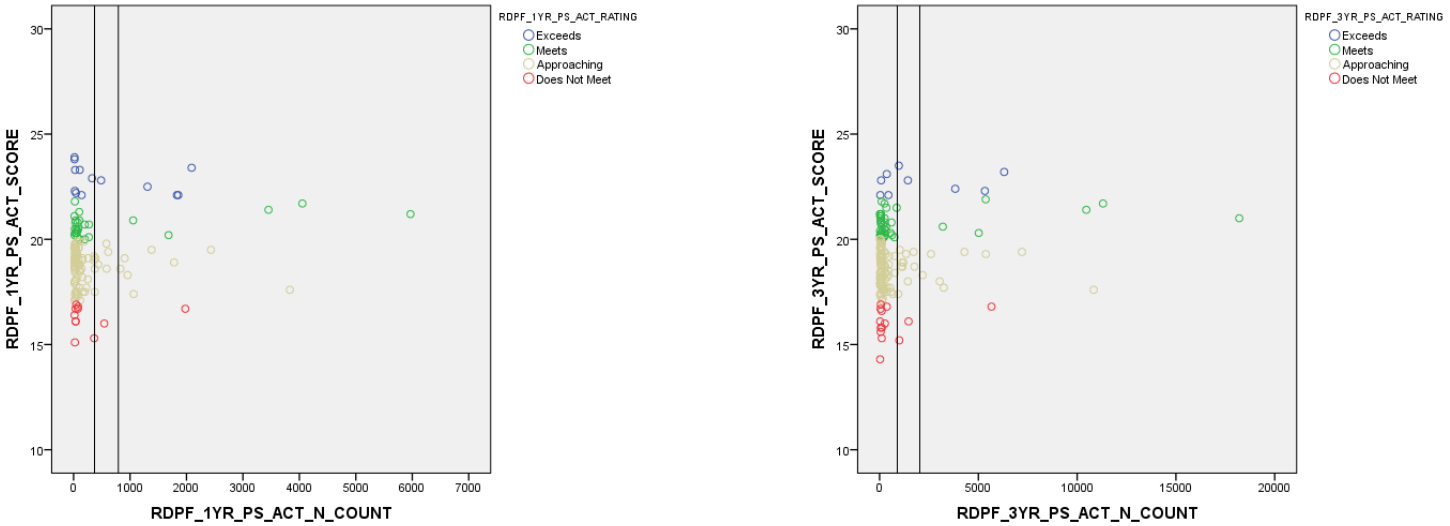
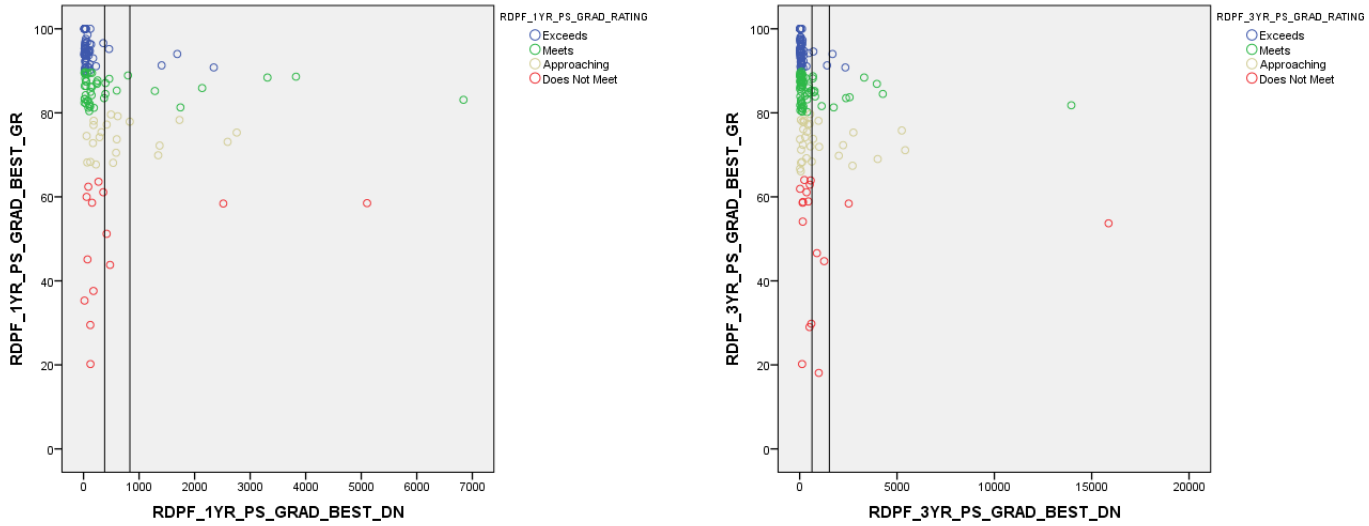


Figure 9. Graduation rates (best) achieved at the district level on one and three year frameworks



Once again, the DPF based plots presented under Figures 8 and 9 highlight the fact that smaller entities are likely to display more variable patterns (i.e. large range) of performance relative to

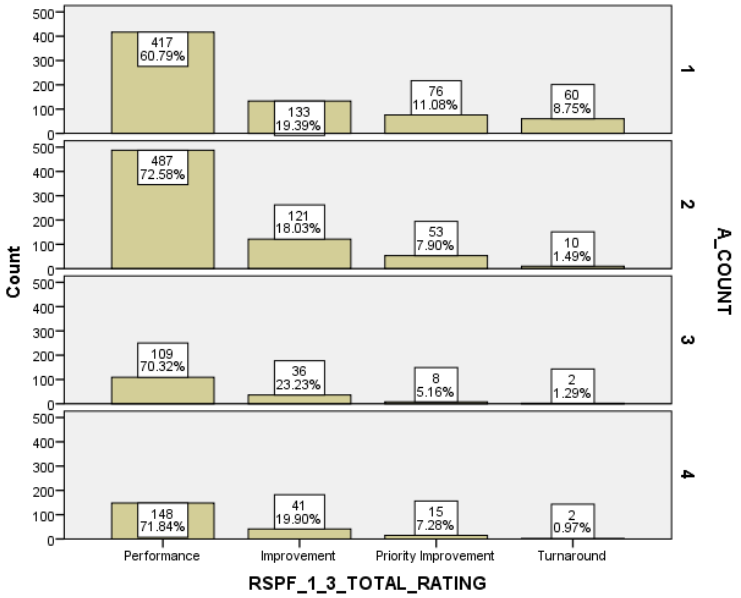
large entities and that these patterns are not simply a result of employing the CGM to evaluate performance across all districts, but also emerge across different types of performance metrics.

School and District Size Relative to Final Overall Ratings

Considering that smaller schools or districts are found in larger numbers each year at the tails of the performance distribution on both growth and proficiency based metrics embedded within the framework, these smaller entities would also likely show up in larger numbers at the low and highest end of the overall ratings distribution prior to any ratings adjustments made for reconsideration cases. Since the final ratings on the DPFs and SPFs are based on a composite index that identifies a set percentage of the lowest performing schools in the turnaround and priority improvement categories, smaller schools and districts are more likely to emerge in the turnaround and priority improvement categories relative to larger schools and districts. Additionally, for the DPFs, since the distinction category is also awarded to schools at the highest end of the performance distribution, smaller districts are also more likely to be found in this category relative to the larger districts. This does not, however mean that large districts cannot attain this status, but rather, compared to small districts, large districts must move many more students in order to achieve this designation.

Figures 10 and 11 present the performance ratings achieved by all schools on the SPFs and all districts on DPFs to n size for the most recent 2012 year. Final SPF performance ratings outcomes viewed relative to n size for other school years are located in Attachment 3. In Figure 10, group 1 represents schools with up to 182 students, group 2 represents schools with 183 – 395 students, group 3 represents schools with 396 – 603 students, and, group 4 represents schools with more than 604 students.

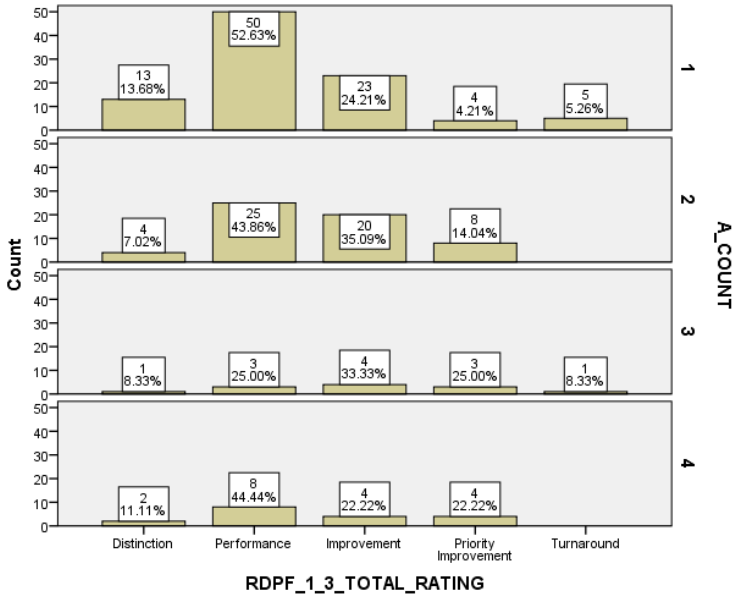
Figure10. Final ratings for all schools by n size on the SPF



In Figure 11, 1 represents districts with up to 350 students, 2 represents district ranging from 351 – 3,000 students, 3 represents districts ranging from 3001-7,000 students and 4 represents districts with more than 7001 students. In each year reviewed (see Attachment 4 to review DPF

final ratings for other years), the maximum number of students exceeded 49,000 and the mean was located at approximately 7,000 students. Therefore categories 1 and 2 capture the majority of districts since as indicated by the scatter plots in Attachment 2, most districts are located well below the mean of the distribution.

Figure11. Final ratings for all districts by n size on the DPF

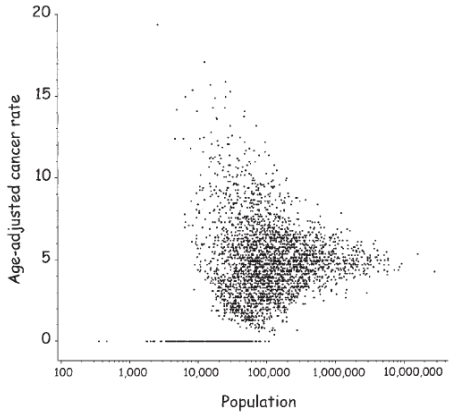


In Figures 10 and 11, and in the results shared in Attachments 3 and 4, the distributions presented reveal that smaller schools or districts (1s and 2s) are recognized as "Accredited with Distinction" in larger proportions than larger schools or districts (3s and 4s). Additionally, smaller schools and districts also appear in larger proportions in the priority improvement and turnaround categories relative to schools and districts classified as 3 or 4.

Discussion

Wainer (2009) notes in his paper, that this commonly observed phenomenon with small n sizes has historically led many policymakers to make incorrect assumptions about the characteristics of smaller institutions across a number of fields. For example, the scatter plot in Figure 11 depicts an illustration of age-adjusted kidney cancer rates relative to a county's population size. The shape of the plot in Figure 11 resembles the funnel shape of the scatter plots found in Attachments 1 and 2 and in all of the scatter plots depicted throughout this document. According to Wainer, at the higher end of the cancer rate distribution, one case of cancer had the effect of disproportionately boosting the cancer rate to high levels for some of the small counties captured in the plot. And at the lower end of the cancer rate distribution, the presence of a larger proportion of small counties appearing at lower levels relative to higher population regions led to false statements made to the public that individuals are less susceptible to cancer in small rural areas.

Figure 11. Age-adjusted kidney cancer rates as a function of county population size



Source: Wainer (2009)

Concerns articulated about the frameworks should not be about whether the growth metrics are inherently "biased" against small or large institutions, but rather should be focused on whether all of the evidence gathered in the framework sufficiently allow for well supported inferences to be made about academic performance taking place over time in smaller institutions.

For larger institutions, considering that these schools and districts benefit from having many more data points to improve the quality of the academic performance signals being estimated or observed, these metrics are likely providing good signals of how students are performing on average at these institutions. In other words, these larger entities appear to be populated with students who perform, on average, in a manner that is similar to the larger population and are likely to have achievement or SGP distributions that resemble the state. In a few cases where districts have argued that large schools suffer from poor performance based on evidence gleaned from other indicators outside of the frameworks, CDE has already implemented the policy of ceding the lower rating decision to the district. Additionally, in the few specific cases where academic performance at larger institutions diverges notably from that of the general population, this finding would provide strong evidence that the institution is either doing an exemplary job of moving large groups of students to higher academic performance levels or is exhibiting alarmingly low levels of academic performance and justifiably requires support.

The primary concern, however, remains to ensure adequate evidence is represented on the framework to warrant a "turnaround" or "priority improvement" rating for smaller entities.

Considerations for Evaluating Small Entities

One check for the quality of evidence reviewed is to evaluate the year to year growth and achievement correlations of these schools. Kane and Staiger (2002, 2008) note that these year to year correlations may serve as reliability estimates that can inform stakeholders whether the signals suffer from a large degree of intertemporal volatility or are fairly stable over time. Should the signals for these small schools appear to be fairly stable, this finding does not mean that the performance signals are accurate, but does imply that the signals are at least consistent and not prone to major fluctuations over time. However, if the correlations are found to be very low (e.g., at or below .3), these findings would require further investigations to identify possible

reasons for these large shifts in performance (e.g., the population of test-takers have changed considerably from one year to the next). Further investigation of the year to year correlations will be provided in the third analysis of this series.

A possible solution to explore for improving the quality and strength of the performance signals for small entities is to ensure that only pooled data (i.e. the three year frameworks) are used and that the n size minimum is reconsidered. A recommendation on a new minimum n size using pooled data would need to be informed by impact studies conducted by CDE to identify how many schools would be excluded based on the set threshold. In the absence of pooled data, if a small entity is flagged for turnaround or priority improvement status, CDE may want to consider an automatic review of additional performance information to be supplied by the institution to ensure that this judgment is supported by additional evidence. This automatic review process may follow the current processes and procedures used to evaluate additional evidence for schools and districts seeking a rating reconsideration from the state. In some cases, the evidence from the frameworks may be deemed sufficient to warrant a low rating. For example, if growth appears to be systematically low across two years on the three year framework, this finding provides compelling evidence that the school or district is not doing enough to increase or improve student academic performance.

Considerations for Small Entities with no Growth Data

Another clear disadvantage that smaller institutions have relative to larger institutions is that they are less likely to be evaluated on growth due to not meeting the minimum n size requirement. That is, a number of these schools or districts are only evaluated on status based metrics and considering that these metrics are known to be highly associated with challenging factors such as poverty, the extent to which an entirely status based framework can fairly evaluate academic performance at those institutions is questionable.

For those entities not meeting the minimum n size on any metric (no performance rating reported) or being evaluated solely on the basis of available status based metrics, CDE may need to consider an alternative process for evaluating performance at those sites. In 2012, there were 12 schools with no data available, and 30 schools with only academic achievement reported. In the same year, there were two school districts with only academic achievement reported. CDE may want to consider engaging these institutions in a process similar to the one mapped out by the state for having schools and districts use the data in the SPFs/DPFs to identify appropriate targets for academic growth in their Unified Improvement Plan. That is, these institutions would set targets that are rigorous but attainable, collect evidence of progress monitoring over the school year to demonstrate how students are moving toward meeting those targets in all grade levels, and be evaluated by CDE or peer reviewers on the extent to which these institutions have met or made considerable progress toward meeting these objectives. Moving in this target setting path would allow for a greater number of sources and evidence to be considered in evaluating performance at these institutions, but would also require CDE to: develop clear and specific guidelines in consultation with the taskforce and potentially other stakeholder groups on what evidence may be considered admissible under this process; and, clearly specify the criteria and develop any rubrics to be used to evaluate the evidence presented by these institutions.

Considerations for "Distinction"

Unlike the SPFs, the DPFs were also designed to reward and recognize districts with exemplary achievement and growth performance. Although the paper shared with CDE by the Poudre School District correctly notes that large institutions are not likely to attain "distinction", their use of the term "biased" is moot. Rather, compared to smaller districts, large districts face a difficult task of moving many more students to higher levels of performance to show exemplary performance on any given metric embedded in the framework. In the same way that smaller institutions being flagged for "turnaround" or "priority improvement" should be checked, smaller entities that reach "distinction" should also be re-evaluated to ensure that sufficient data are being used to evaluate whether this judgment is warranted. That is, pooled data from the three year frameworks should be used for smaller institutions to ensure that the best available information is used to inform these judgments.

An alternative approach that CDE may want to consider for this distinction category is to extend the set of criteria outside of the state assessment focused metrics specified in the frameworks. That is, districts with "performance" ratings may be considered for "distinction" if they demonstrate excellence in a combination of other areas such as:

- Sustained low levels of truancy each year;
- Increased completion rates of FAFSA applications;
- Reduced remediation rates in higher education; and,
- Increased student participation in higher level courses and acquisition of post-secondary credit.

Allowing for the inclusion of a select group of additional indicators to warrant a distinction status may potentially bolster opportunities for larger districts exhibiting strong academic performance to attain this status. These indicators should only be considered and selected if they clearly align with and support state goals to prepare students for post-secondary and workforce readiness.

Considerations for Developing Separate Frameworks based on Size

One idea that has been advanced by some district stakeholders is to develop a separate set of frameworks for schools and districts based on n size. However, as seen in Attachments 3 and 4, the majority of districts and schools evaluated by the frameworks consist of smaller entities, defined in this paper as being located below the average school or district size. Considering that the cuts set for growth and for all other metrics were set using the entire distribution of schools and districts on the framework, this implies that the frameworks are largely designed to differentiate performance among these smaller entities. The value-added of developing a new framework for differentiating performance among larger entities that share similar performance profiles (unless these are alternative education campuses) is unclear. For these larger entities, it may be more useful to differentiate performance by drawing upon, as described in the previous section, an additional set of indicators that reside outside of the points-based framework.

Conclusion

As CDE and stakeholder groups continue with the iterative process to improve upon the performance frameworks, it is important to keep in mind that the primary purpose of these frameworks is to ensure that districts and schools suffering from systematically low levels of performance are flagged for support. Although smaller entities will be represented in higher numbers each year for support, the signals for these schools should always be checked to ensure that the best available information about the school's performance is considered prior to confirming those ratings. As noted earlier, CDE has already taken considerable steps to build in checks into the system by allowing schools and districts the option to submit reconsiderations and evidence supporting a case for higher ratings and has also given districts the right to lower ratings for those schools they deem to exhibit lower performance on other metrics that fall outside of those used in the SPFs. However, this paper recommends that CDE take the additional step to ensure that low ratings for small schools exhibiting volatile performance movements over two years are automatically verified in order to ensure that these ratings are informed by pooled data and in the absence of pooled data, are informed by a process similar to the current reconsideration process. Additionally, this paper recommends the consideration of additional indicators outside of the framework to provide other opportunities for larger districts to demonstrate exemplary performance.

References

Briggs, D. C. & Betebenner, D. (2009) Is Growth in Student Achievement Scale Dependent? Paper presented at the invited symposium "Measuring and Evaluating Changes in Student Achievement: A Conversation about Technical and Conceptual Issues" at the annual meeting of the National Council for Measurement in Education, San Diego, CA, April 14, 2009

Kane, T. J., & Staiger, D. O. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives*, *16*(4), 91-114.

Kane, T., & Staiger, D. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. *NBER working paper*. Retrieved from http://www.nber.org/papers/w14607.

Wainer, H. (2009). *Picturing the Uncertain World: How to Understand, Communicate, and Control Uncertainty through Graphical Display*. NJ: Princeton University Press.

Wright, P. (2010). An Investigation of Two Nonparametric Regression Models for Value-Added Assessment in Education. SAS White Paper. http://www.sas.com/resources/whitepaper/wp_16975.pdf

Attachments

Attachment 1: 1-year and 3-year SPF reading and math growth charts by n size and level

Attachment 2: 1-year and 3-year DPF reading and math growth charts by n size and level

Attachment 3: Final ratings on the SPFs achieved by all schools and by year

Attachment 4: Final ratings on the DPFs achieved by all districts and by year