TPE Communication Bulletin #29: January, 2015

Office of Teacher and Principal Evaluation

Dave Volrath: *Planning and Development*
Tom DeHart: *Aspiring & Promising Principals*
Liz Neal: *Institutes of Higher Education*
Ilene Swirnow: *Professional Development;Executive Officers and Principals*
Joe Freed: *Professional Development; Teachers and Principals*
Frank Stetson: *Professional Development; Teachers and Principals*
Ben Feldman: *Technical Assistance*
Laura Motel: *Communications*
Teri Windley: *Administrative Assistance*

Special Edition: 2014 Educator Effectiveness Ratings

The integration of Teacher and Principal Evaluation with MSDE organizational and technology structures

On October 24, 2014 Maryland released its first-ever statewide accounting of Teacher and Principal Effectiveness (TPE) Ratings. In accordance with agreements in Maryland's Race To the Top Grant, twenty-two Local Education Agencies (LEAs) provided ratings of Highly Effective, Effective, or Ineffective for every eligible teacher and principal in the state. In compliance with ESEA Principle 3, the remaining two LEAs will provide effectiveness ratings for their teachers at the end of the 2015-2016 school year. TPE ratings information were presented in public session to the Maryland State Board of Education and are posted to the Maryland State Department of Education (MSDE) website.

The report presents the accomplishment of twenty-two approved local models, a collection of approximately a half million data points. "Approved" models had to demonstrate the intention of the Education Reform Act of 2010 to balance measurement of Professional Practice with quantifiable evidence of Student Growth. Additional parameters mandated certain minimum domains or outcomes for teacher and principal Professional Practice and the use of multiple measures to assess Student Growth.

This work was executed between January 2013 and June 2014 and would have been impossible without the commitment and collaboration of LEAs, Superintendents, Maryland State Education Association, Baltimore Teachers' Union, and the MSDE TPE Team. It is because of the completion of these promised deliverables that this work has been approved by United States Department of Education for a fifth year no-cost extension of Race to the Top (RTTT) support.

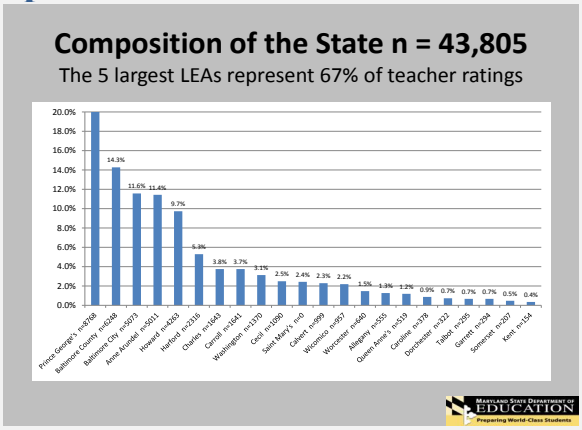
The successful data collection reflects close collaboration with LEAs in the design of the reporting methodology, persistence in the authentication of the accuracy of the data, and transparency in the presentation of data by the districts and by the state. As we trumpet these programmatic accomplishments, it is important to note that while interesting trends are evidenced, proclamations of degrees success will take months if not years. This information is intended to generate the questions that must now be answered by the state and the LEAs to make sense of the data, to validate the work, and to point towards the next direction.

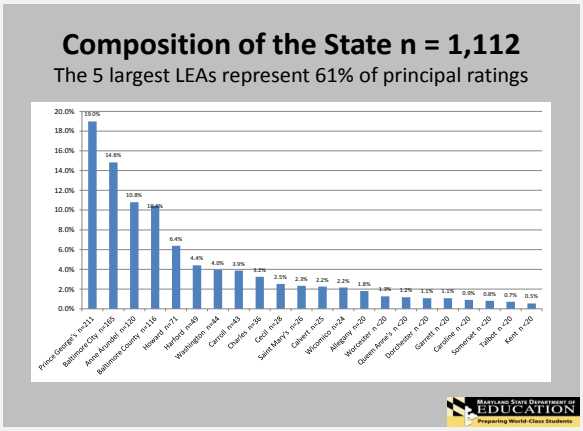
- The present report is a *descriptive* analysis of 43,805 teacher and 1,112 principal ratings provided by all 22 RTTT LEAs.
- The *inferential* statistical analysis is being conducted by MACC@WestEd.
  - This independent report will examine the performance of the models and their components.
  - This report is expected toward the close of the calendar year.
- LEAs should conduct independent analyses that may replicate the State's approach.
- Throughout this report, MSDE offers LEAs direct suggestions, indicated in red.

By spring 2015, LEAs will be able to refine their models. LEAs know their people and their models best, a parallel analysis by LEAS of similar dis-aggregations and local interests needs to be undertaken. From this effort will arise refinements to local evaluation resulting from discoveries from this first year of full implementation will inform state and local evaluation for the 2015-2016 school year

The complete PowerPoint presentation to the Board, and also detail LEA by School files and LEA summary files for teachers and for principals are online: LEA/School Teacher-Principal Evaluations

Composition of the Statewide Data



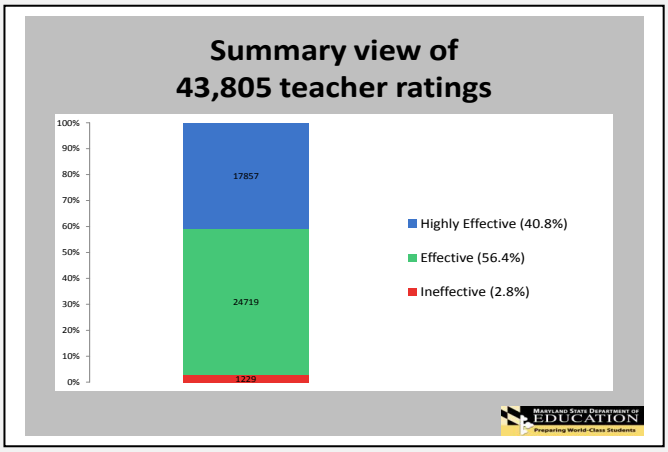


The first two slides illustrate the composition of the 43,805 teacher and 1,112 principal ratings. The five largest LEAs, of which Prince George's and Baltimore County are the largest, account for nearly two thirds of all ratings. In all subsequent slides but one, blue indicates "Highly Effective," green indicates "Effective," and red indicates "Ineffective."

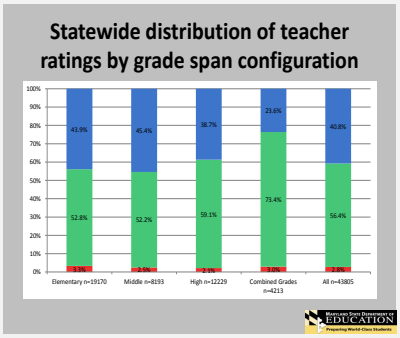
Analysis of 2013-2014 Effectiveness Ratings and Component Evaluation Measures

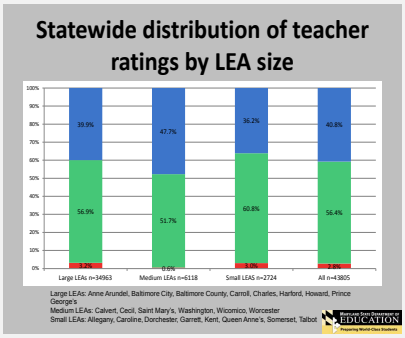
State Teacher Ratings

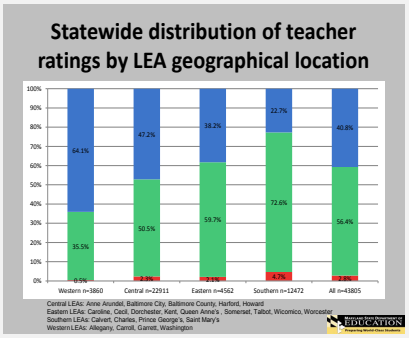
Of the 43,805 teacher ratings provided by LEAs, 42,576 were rated as effective or highly effective. The remaining 1,229 or 2.8% were rated as ineffective. This proportion rated as ineffective is more than twice the percentage reported as "unsatisfactory" in the most recent published report of teacher performance statistics.



By Grade Level, Size, and Geographical Location

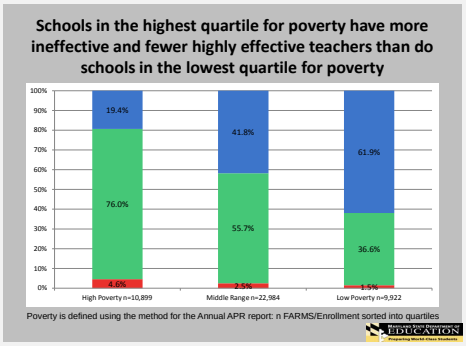


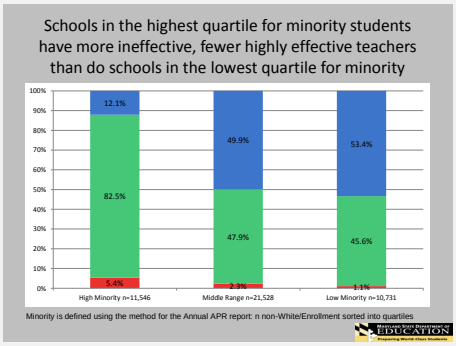




The distribution of teacher ratings is consistent between elementary and middle schools. Although there were fewer teachers rated as ineffective in high schools, there were also fewer teachers rated as highly effective in high schools. "Combined grade" schools had the smallest proportion of highly effective teachers. These schools do not fall into the standard K-5/6-8/9-12 configurations, and in some LEAs may represent special programs and special populations. When LEAs were examined by size, the medium sized LEAs had the most favorable proportions of effective/highly effective ratings. The most visible patterns are observed by geographical regions, with the western LEAs having the highest proportions of effective and highly effective teacher ratings, and the southern LEAs having the smallest proportion of high effective and the preponderance of ineffective teacher ratings. Implications for LEAs: LEAs should replicate these analyses within their districts to identify variation of teacher ratings among school programs, school sizes, or location within the jurisdiction.

Poverty and Minority Impact

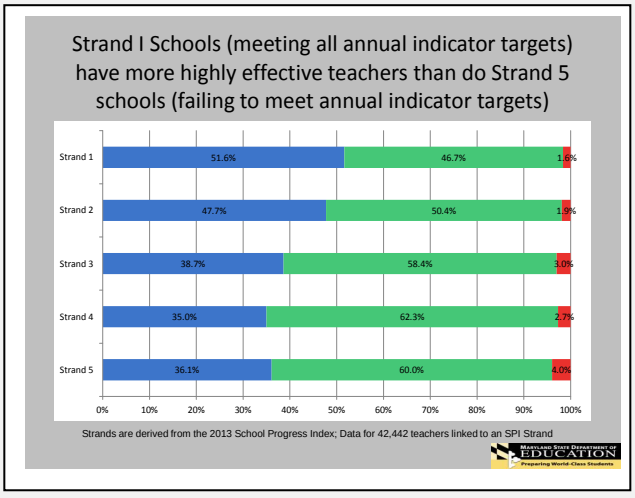




The proportion of highly effective and ineffective ratings stratifies with the official indicators for poverty and minority used for federal reporting. To create these indicators, the numerator includes those students identified as FARMS-eligible or of a minority race-ethnicity code; the denominator includes all pupils-in-membership. Using this proportion, schools are arrayed statewide and sorted into quartiles. In the above graphs, "high" and "low" reflect the first and fourth quartiles respectively. Although there has been a "sense" or perception that teacher effectiveness might be related to whole-school demographics, these analyses confirm this association. In Maryland, "poverty" is a more widely distributed attribute, encompassing rural poverty as well as urban poverty. Minority effects are more closely clustered in specific areas, and for the highest concentrated quartile, the correlation to highly effective and ineffective is most evident. Implications for LEAs: These are important and concerning findings and LEAs should replicate these analyses using their own schools as the unit of analysis.

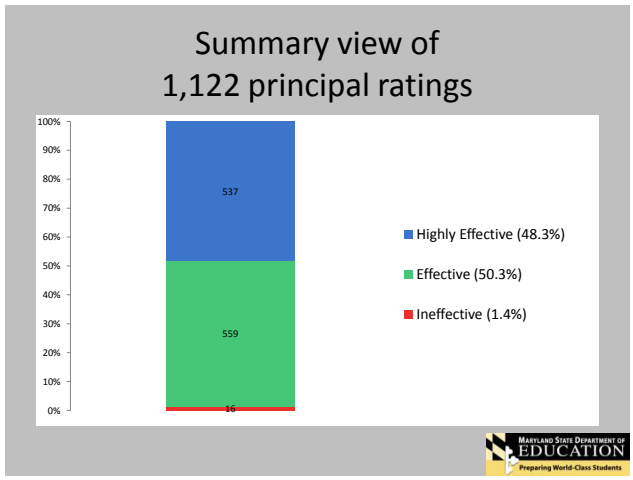
School Performance

During the transition period from MSA to PARCC assessments, school performance statistics are less available that in prior years and will continue to be in immediate coming years. Moreover, the ESEA waiver allowed the spring 2013 School Progress Indicator, or SPI, to remain in place for 2014. The SPI sorts schools into performance strands from one to five, one indicating a school that meets all of its indicator targets and five indicating a school that fails to meet its targets. Teacher ratings stratify according to SPI strands with Strand 1 schools having the highest percentage of highly effective teachers and the lowest percentage of ineffective teachers. Implications for LEAs: LEAs should ascertain if there is variation in teacher ratings against SPI ratings, as such findings are even more meaningful at the LEA level than at the State level. One salient question LEAs should ask themselves is whether schools are being "engineered" for success when assigning staff.

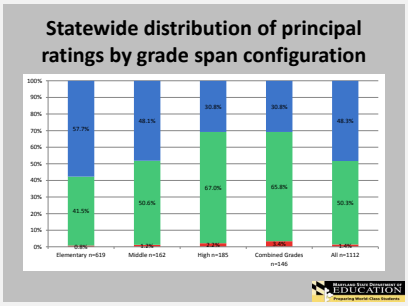


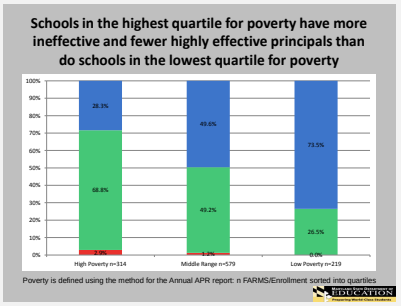
State Principal Ratings

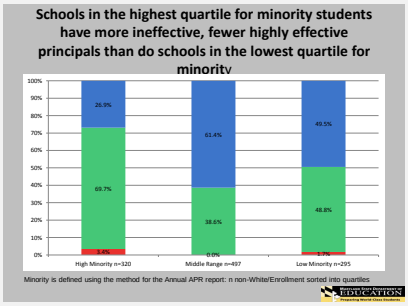
Of the 1,122 principal ratings provided by LEAs, 1,106 are highly effective or effective, split almost evenly. This is not surprising as the principalship is an at-will position, and superintendents may evidence willingness to make strategic assignments to benefit needs of school communities.



Grade Levels, Poverty and Minority Impact



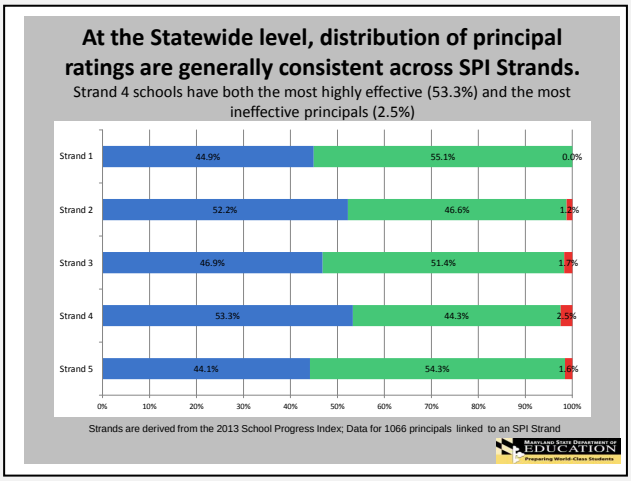




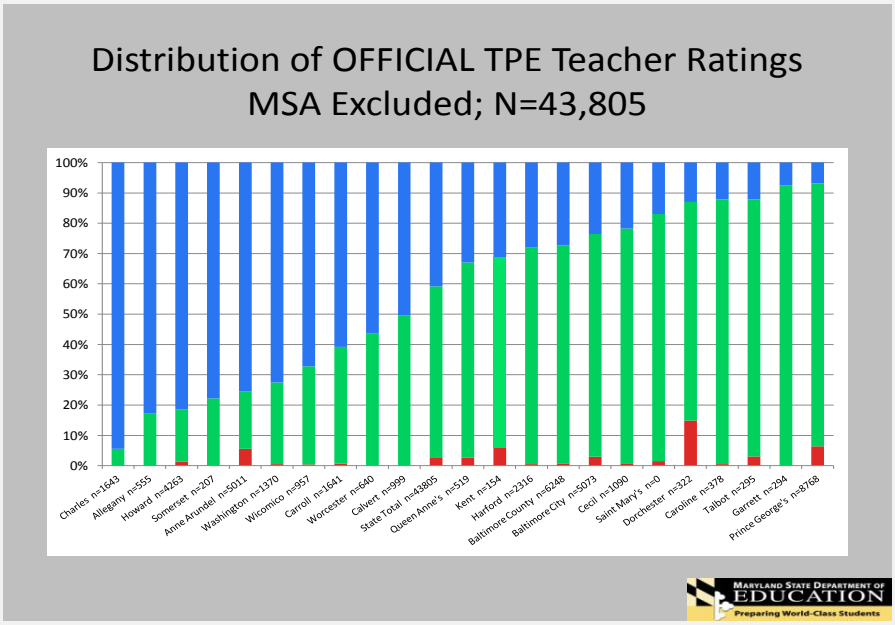
The performance distributions of principal ratings for grade levels, for concentration of poverty, and for concentration of minority populations mirror those distributions seen in teacher ratings. The decreasing proportion of highly effective principal ratings in secondary and combined grade programs may reflect structural challenges associated with these larger, more complex programs. The stratification of ratings with poverty and minority concentration, observed in teacher ratings, is equally explicit and concerning.

School Performance

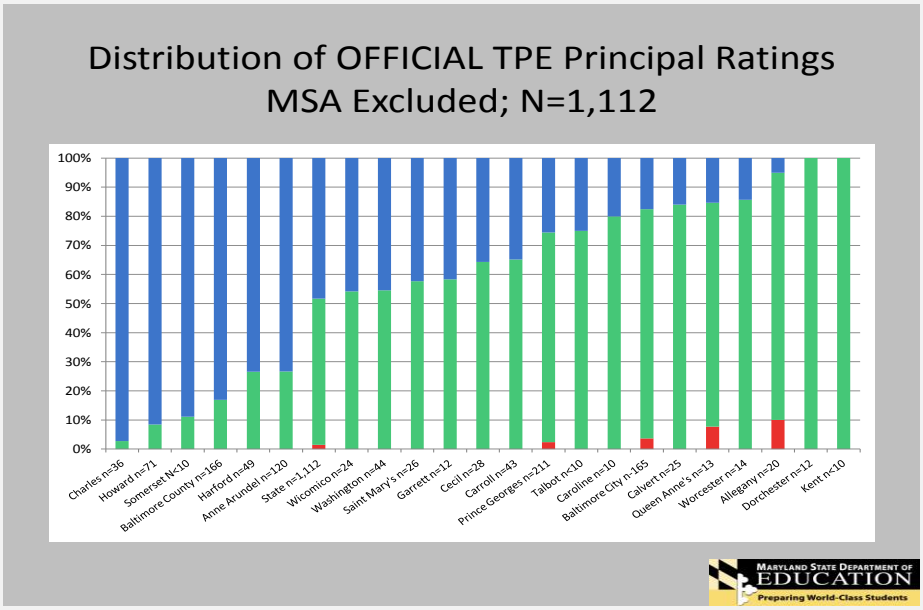
Principal ratings do not stratify cleanly with SPI strands as do teacher ratings. Moreover, the observation that ratings tend to look weaker in Strand 3 and 4 schools rather than in Strand 5 schools may suggest the attention that superintendents and executive officers bring to schools facing the strongest ESEA sanctions and a willingness to place more effective leaders in situations requiring turn-about skills.



LEA Rating Distributions

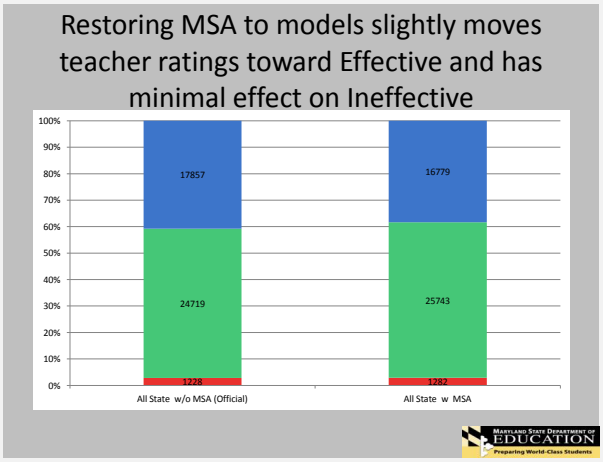


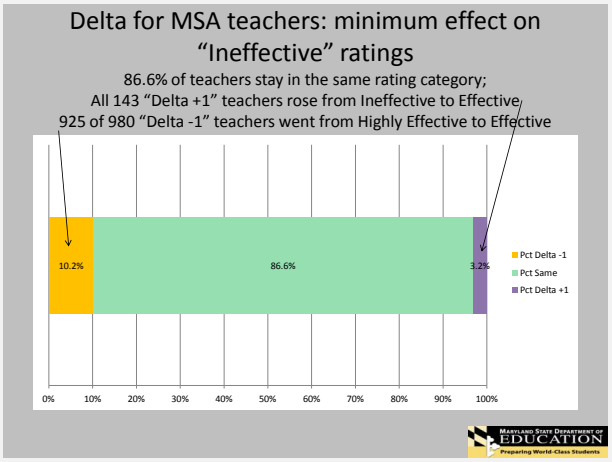
The above figure arrays the 22 participating RTTT LEA teacher ratings sorting by the percentage of highly effective. The State Average falls to the left of the center of the distribution. At this moment in the history of the project, these ratings must be viewed with caution. Three things are certainly operating: 1) there are genuine differences in the ability of staff; 2) the models are functioning differently, and most important, 3) cut scores distinguishing among rating levels reflect difference levels of precision. For example, the LEAs at the extreme left may have generally superior staff; it is also possible that they may have set cut scores low in the first full consequential year. Implications for LEAs: LEAs should scrutinize how their teacher ratings fall within the array displayed. LEAs need to critique the extent to which their TPE models are successfully discriminating among levels of performance. LEAs need to carefully examine their own staff at transition points and submit their findings to local expert judgment.



Principal data is arrayed the same way as teacher data with a descending sort on highly effective. Although not a "carbon copy," LEAs generally occupy the left or right side of the graphic for principals much as they did for teachers. Implications for LEAs: The same issues and questions remain on the table: Are staff genuinely different in ability? Is the local model discerning? Have cut scores been set with precision? To this must be added a critical question: Are LEAs having the difficult conversations that must occur between staff—teachers or school leaders—and their evaluators if TPE is to be an effective driver of improvement through Professional Development?

MSA Impact





The role played by the State Assessment has been one of the most debated aspects of TPE since the passage of the Education Reform Act of 2010 and the winning of the RTTT Grant. The two figures above show this controversy has not played out as many argued it might. For context, the original MSDE parameter required that the Maryland School Assessments (MSAs), where they existed, needed to represent 20 percentage points of the total evaluation. Under the flexibility provided by USDE in fall 2013, Maryland was allowed to set aside the MSA, although the equal split between Practice and Growth had to remain intact. All LEAs ran their approved models WITH and WITHOUT the MSA. The version without yielded the "official" rating of record.

First, it should be noted that restoring the MSA at 20% value had almost no effect on the overall distribution of teacher ratings. The effect, such as it is, is to better center the data within the Effective rating range. Statewide, only 54 additional teachers were rated as Ineffective out of a population of nearly 44 thousand teachers (0.1%). However, when these data are examined, not in the aggregate but as an actual effect on individuals, the results are more interesting. To visualize this relationship a "delta" variable was created. A positive delta value indicated a rise in performance rating and vice versa. A delta value of zero meant the rating was unchanged with or without addition of the MSA. Nearly 87% of all MSA teachers were unchanged, but of the 143 teachers who earned a +1 delta, every case represented a rise from Ineffective to Effective. For these teachers, the MSA always boosted their rating. Of the 980 teachers who earned a -1 delta, 925 or 94.4% of them fell from Highly Effective to Effective, demonstrating the centering effect mentioned above.

Implications for LEAs: LEAs need to carve out these staff and examine these records closely. Did the MSA introduce objectivity and reduce subjectivity in the assignment of ratings? Particularly, for those teachers who were raised from ineffective, were concerns revealed in the consistency and robustness with which the Professional Practice half of the evaluation model was applied?

Of great interest, is data related to the investigation of methodologies for the setting of cut scores for Highly Effective and Ineffective ratings. This is a discussion that has been on-going nationally for the past five years and was typically addressed with arbitrary or intuitive determinations. The current data may allow Maryland to conduct stress testing that demonstrates the levels at which the cut scores begin to initiate movement and to validate the efficacy of ratings at different cut score settings. Implications for LEAs: The State highly recommends that LEAs contribute to this exploration by conducting similar scenario-driven investigations using actual local effectiveness rating data. These findings could be critical to conversations about the design, performance, and sustainability of model designs in when Student Growth Measures reenter the conversation in June of 2016.

A strategic plan, which includes a quality control process and addresses the primary interests of Student Learning Objectives, Communications, and Sustainability

Discussion, Recommendations and Next Steps

The first and most important statement to reiterate is that Maryland successfully rose to the challenge set by the Reform Act and RTTT: to rate every eligible teacher and principal using an approved model which balanced consistent Professional Practice domains or outcomes with quantifiable measures of Student Growth. This accomplish has stumped most of the nation, and despite many hours of debate and controversy, Maryland achieved this using local models that evidenced a surprising degree of uniformity of design. Every LEA brought this task to a landing, and every LEA reported data in a normalized form that allowed for comparison across LEAs. The magnitude of this to completion, arguably the most consequential of all RTTT, cannot be overstated.

However, the data and the models have not matured, and moreover, this work has been conducted during a period when the MSAs sunset, a new assessment (PARCC) is not yet in place, and the curriculum changed. No one expected the TPE ratings to represent a completed settled piece of work. They are not, but the first year's work is nevertheless very revealing. Some of the take-away observations from the data includes:

- Teachers and principals in different LEAs fared differently, especially in distinguishing Highly Effective from Effective ratings.
- Although some structural variables such as grade span, LEA size, and geographical location produce some observable differences, these differences are dwarfed by the effects of concentrating poverty and minority populations of students. Poverty, in particular, is a statewide story, and whether one considers urban poverty or rural poverty, it is a concern seen among principals' ratings as well as among teachers' ratings.
- Although incorporating state assessments had been a hot issue since the project's inception, the assessments appear to have centered the data and more often benefited teachers rather than harming them.
- While state-level analyses suggest contours, meaning must be sought at the LEA level, at the dynamic within individual schools.
- The strongly positive disposition of the data suggests that LEA leadership continues to approach difficult conversations with a degree of trepidation.

Recommendations

- ✓ LEAs should replicate the State's analyses as a point of departure. LEAs should examine differences among their elementary, middle, high, and special configuration schools. LEAs should ascertain if small schools are performing differently than large ones, especially for differences among similar grade bands, e.g., small middle schools versus large middle schools. LEAs need to ascertain if there are geographical effects within a county, especially larger counties.
- ✓ LEAs must take a hard look at concentrations of poverty and minority populations and the caliber of staff assigned to those schools. LEAs must question whether results are real or artifacts of perception and expectation.
- ✓ LEAs should sort all of their teachers by total accrued points and carefully study the rankings of staff—informed by local expert judgment—and most particularly at the transition from one rating level to another. The State is convinced that while some LEAs will find that the scoring and rating categories are discrete, others will find that scores and categories are not aligned. Wherever this is the case, executive officers and principals must have probing conversations to learn how evaluators arrived at ratings. LEAs should receive this admonition with attention and seriousness.

The next major event will be the release of the MACC@WestEd independent analysis of model performance. With these results in hand, and if coupled with LEA analyses responding the State's challenges offered throughout this paper in red, LEAs will be equipped to engage with their communities in early spring to make important refinements to their models. Such refinements may include adjusting the weight of component parts, a more rigorous address to assessing SLOs, and most particularly, using empirical evidence to inform the setting of cut scores.

Again and again the State references the application of "local expert judgment." Every piece of quantitative data has a qualitative story beneath it. LEAs must enter into a vigorous self-study of their ratings, challenging the appropriateness of every decision. LEAs must prepare themselves for the period when holding a "developing" teacher as effective will no longer be an acceptable option. Most significantly, Effective as a rating category must convey rigor; it cannot be used to conceal unacceptable weaknesses.

Maryland's success with TPE is largely predicated on its focus on improvement: using evaluative data to elevate the quality of the instructional cadre through strategically informed Professional Development. Teachers throughout the state have reported in surveys and focus groups that their reflections are deeper and their professional conversations have been more purposeful. Going forward, these conversations must be more authentic than ever. If the outcome is to be about professional development and instructional improvement, rather than compliance and accountability, Leadership must be willing to lay out internal data school by school and case by case. If there is to be credibility in our efforts to achieve better outcomes for students and their families, then LEAs must become experts in understanding the performance and veracity of their models and commitment to the fidelity and transparent application of those models.