Research Article

Problems with Percentiles: Student Growth Scores in New York's Teacher Evaluation System

Drew Patrick, MEd
Assistant Superintendent, Curriculum & Instruction
Bedford Central School District
Mount Kisco, NY

Abstract

New York State has used the Growth Model for Educator Evaluation ratings since the 2011-2012 school year. Since that time, student growth percentiles have been used as the basis for teacher and principal ratings. While a great deal has been written about the use of student test scores to measures educator effectiveness, less attention has been paid to how value added models have played out in schools, school districts, and states since their widespread adoption associated with Race to the Top. This study employs univariate and multivariate statistical procedures to examine model results at the student level in one district, and across districts, and identifies problems associated with the model. Policy implications and recommendations are discussed.

Key Words

Growth models, value-added modeling, student growth percentiles, educator evaluation ratings, evaluation policy

Introduction

The use of test scores to evaluate teachers and principals has increased tremendously during the Race to the Top era (Baker, Oluwole, & Green, 2013). Generally referred to as *value-added modeling* (VAM), the technique relies on complex statistical models to predict future student test scores based on prior scores and various other demographic and school-related factors. Teachers and principals of students who beat their predictions are considered to have "added value", or contributed substantially to student learning, relative to teachers whose students miss their predicted scores.

In many systems, teachers are then assigned to a rating category (i.e., "ineffective" or "highly effective"). Not surprisingly, lines have been drawn and a debate is underway between proponents of VAM and those who argue against its utility for gauging educator effectiveness (Goldhaber, 2015; Holloway-Libell & Amrein-Beardsley, 2015). While it is important to understand the arguments for and against, briefly outlined in the next section, there remains a substantial dearth of information about the performance of state-or district-specific VAMs over time. There is a clear and present need for a determination as to whether or not these models are capable of producing the results intended by the policy makers who adopted them.

The purpose of this exploratory study was to gauge the extent to which the New York Growth Model for Educator Evaluation provides meaningful student level growth data to inform educator practice. Furthermore, since these student-level data are aggregated at the teacher and school level to make effectiveness determinations, the study also attempted to identify potential problems with their use for this purpose. Overall, the analysis raises concerns about the meaning of student growth percentiles (SGPs), along with questions about year-to-year stability and performance-level bias, such that using these measures to assign a teacher or principal growth score deserves closer examination, and supports the call for a broader and deeper study.

Context of the Problem

The use of student growth for accountability purposes first entered the education policy arena in the context of the school and district-level performance, as opposed to teacher performance (Betebenner, 2011). In 2005, the USDOE gave states opportunities to begin measuring and reporting student growth-toward-proficiency as a strategy to meet AYP (adequate yearly progress) as part of the Growth Model Pilot Program (Hoffer et al., 2011). As the accountability gears kept grinding, the methodologies associated with this (i.e., VAM) were turned toward the classroom (Betebenner, 2011). Since this time, economists and educational researchers have been debating over the use of these models for teacher-level accountability.

Research that favors the use of VAM to make judgments about educator effectiveness generally argue that the potential for good outweighs the negatives, and is constructed around the following ideas (Chetty, Friedman, & Rockoff, 2014a; Chetty, Friedman, & Rockoff, 2014b; Hanushek & Rivkin, 2010; Holloway-Libell & Amrein-Beardsley, 2015; Rockoff & Speroni, 2010; Tyler, Taylor, Kane, & Wooten, 2010):

- Teachers' effectiveness varies as measured by value-added
- Teacher value-added is an educationally and economically meaningful measure
- Teacher effects can be discerned from VAMs in an unbiased manner
- The models and the results they produce can adequately control for non-classroom or teacher effects
- Using teacher value-added improves achievement more than not using it

On the other side, those opposed to using VAM for educator effectiveness decisions argue there is a high risk of unintended negative consequences, including false positives and negatives, narrowing of the curriculum, class roster and student test manipulation.

These arguments are constructed around the following ideas (Baker et al., 2013; E. L. Baker et al., 2010; Ballou & Springer, 2015; Braun, 2015; Darling-Hammond, 2015; McCaffrey, Lockwood, Koretz, & Hamilton, 2003; Rothstein, 2010; Strong, Gargani, & Hacifazlioğlu, 2011):

- Value-added estimates are biased, and are invalid- they do not measure what they purport to measure
- Value-added estimates have unacceptably high error to be used in making high stakes decisions about teachers
- Value-added estimates are unstable over time, limiting their reliability and therefore usefulness
- Value-added estimates are too complex to be understood in meaningful ways by those for whom they are intended (i.e., teachers and school leaders)
- There are underlying biases in the student-level estimation of growth that create problems for aggregating to teacher-level effects
- Even if you identify bad teachers with VAM, the current workforce does not support the idea that low performers can regularly be replaced by higher performers.

Regardless of viewpoint, models that rely on student test scores to make educator effectiveness determinations are in use across the country (Baker et al., 2013). One of the gaps in the literature is a lack of research focused on the value-added models that are currently in place in states and districts. Specifically, it is important to examine how these models have performed over time with respect to their ability to predict student performance in a meaningful way, and therefore contribute toward an understanding of teacher influence on that performance.

New York's Growth Model for Educator Evaluation

Student growth on state tests as determined by New York's Growth Model for Educator Evaluation, developed by American Institutes for Research (AIR), has been used over the past four years to generate one of the multiple measures used in deriving an overall teacher score and rating (American Institutes for Research, 2014). This model results in state-provided growth scores (SPGS) for teachers of ELA and mathematics in grades 4-8. This score represents 20% of an overall composite score that also includes locally-determined measures of student growth or achievement (20%) and other measures based on classroom observation (60%).

While only impacting some 15% of classroom teachers, the use of SGPs has implications beyond just 4-8 ELA and mathematics teachers because some school districts elected for the simplicity of applying state-provided scores to all teachers, something allowable under New York's evaluation law (New York State Education Department, 2012).

The model uses grade-specific multiple regression equations to generate predictions for current year test scores, taking into account up to three years of prior tests scores, along with various demographic and other factors. Recently, the reliability of these predictions was called into question in the form of a legal challenge.

In August 2015, oral arguments were heard in a case brought against former NYS Education Commissioner John King by a fourth grade teacher from Long Island. The teacher sought a remedy to the arbitrary and capricious nature of her SPGS, which dropped from 14/20 in 2012-13 to 1/20 in 2013-14. Part of the case centered on the influence a single student's test score had on the teacher's score.

One student received a perfect score on the state test prior to entering the plaintiff's classroom, and the growth model predicted another perfect score in 4th grade. The student ended up getting a total of two questions wrong, which lowered the teacher's score into the ineffective range. The student's score was higher than 99% of all 4th graders state-wide, but the teacher was rated in the bottom 6% in part due to this "failure" (B. Lederman, personal communication, August 12, 2015).

While a decision is pending in this case, the New York State Education Department made a significant policy change in September 2015, creating a process by which, under certain conditions, teachers and principals can appeal their SPGS and have it thrown out (New York State Education Department, 2015b). On its face, this change intimates concern by the Education Department about the ability of the model to produce meaningful results.

The Study

Description

This study aimed to explore the question, *To what extent does the New York Growth Model for Educator Evaluation provide meaningful student level growth data to inform educator practice and gauge effectiveness*? To answer this question, the study relied on an analysis of a region-level (16 school districts), and district-level (1 district) dataset based on the 2015 New York State English language arts (ELA) and mathematics tests.

Each dataset included de-identified student-level data for: test name, current and prior-year (2014) scale score, current year predicted scale score, current and prior-year (2014) performance level, and current and prior-year (2014) percentile rank. Calculated variables included change in performance level (2015-2014) and change in percentile rank (2015-2014), categorized into deciles (0-10 = 10, 11-20 = 20, etc.). The district-level data set also included student growth percentiles, where available, back to 2011-12. It is important to note that growth percentiles are first generated for students in grade four, as that is the first possible year in which students have a prior-year test score, the most important independent (predictor) variable in the growth model.

Accordingly, grade eight is the last year in which student growth percentiles are

calculated. The question was answered using descriptive statistics (frequencies, means) and one-way analysis of variance.

Limitations

The study had several limitations. First, the region-level dataset did not contain SGPs from years prior to 2015, preventing a broader examination of the year-to-year stability of SGPs in the larger sample.

Second, these results, while based on a large N-size of over 4,300 students, may not be generalizable to populations in other parts of the state, in part because the percentage of students on free or reduced lunch is much lower than the state-wide average.

Third, more complex statistical analyses need to be done to further explore the correlations between SGPs year-to- year.

Findings

The analysis began with an examination of ELA results. Descriptive statistics related to the 16-district dataset are presented in Table 1.

The proportion of English language learners (ELL) and students with disabilities is similar to the state-wide average (3% and 8%, respectively; New York State Education Department, 2015a), but the free or reduced lunch percent-age falls 10% shy of the state average. The mean SGP ranges from 47.0 (grade 4) to 51.2 (grade 5), with an overall value of 49.6, very close to the expected mean of the state-wide distribution.

This suggests that sample population, overall, exhibits characteristic student growth behavior, hovering at the mean.

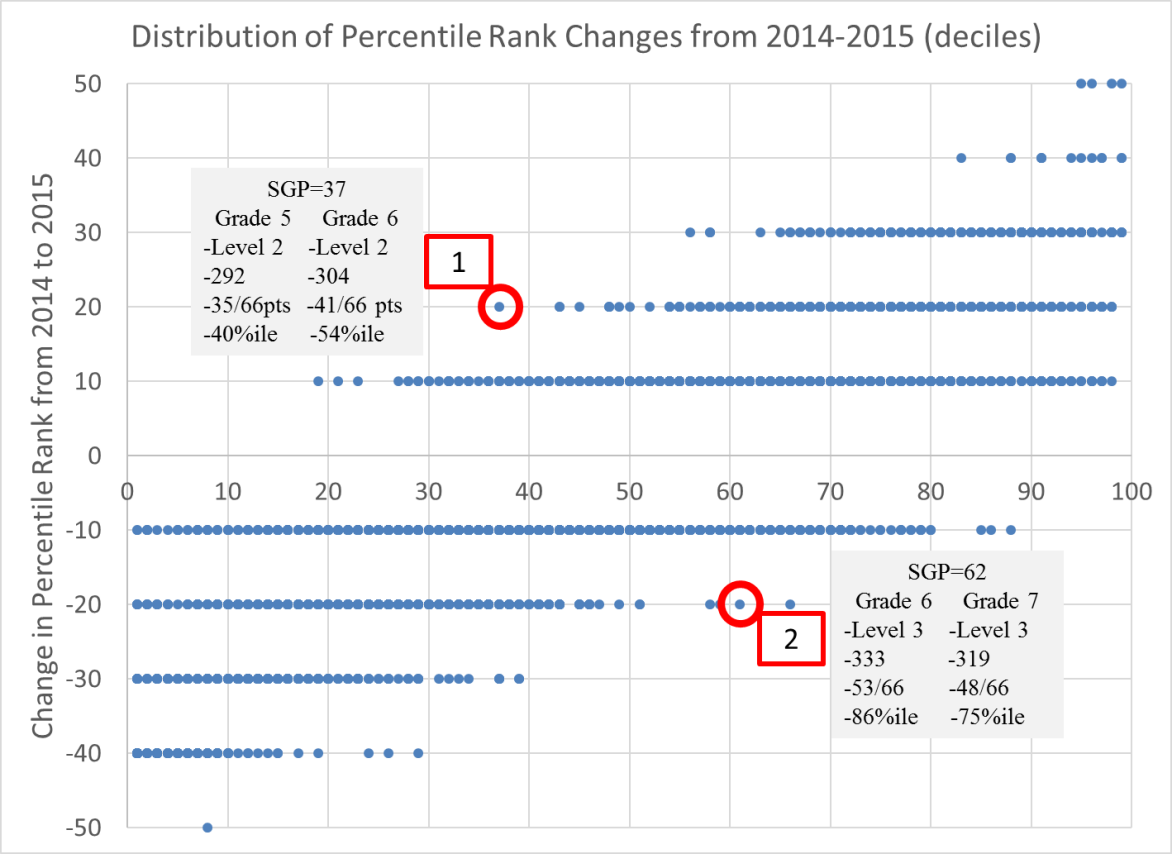
Table 1

Descriptive Statistics for 2015 Region-Level Student ELA Data

Grade	N	Free or Reduced Lunch	Students with Disabilities	English Language Learners	Mean ELA Student Growth Percentile
4	889	145	75	24	47.0
Female	422	64	24	11	47.6
Male	467	81	51	13	46.4
5	905	138	92	25	51.2
Female	428	66	34	6	53.2
Male	477	72	58	19	49.5
6	898	122	86	23	49.5
Female	431	54	38	11	51.9
Male	467	68	48	12	47.2
7	863	116	88	30	50.9
Female	417	46	31	16	53.4
Male	446	70	57	14	48.5
8	812	124	81	21	49.5
Female	425	71	38	13	51.1
Male	387	53	43	8	47.6
Total	4367	645 (15%)	422 (10%)	123 (3%)	49.6

Figure 1 (below) shows more detail by illustrating the distribution of the ELA SGPs grouped according to changes in overall achievement percentile rank for students in the dataset. The percentile rank for a student represents the overall percentage of students state-wide which that student outperformed on the same test in the same year. The change in this achievement measure was calculated by subtracting the 2014 rank from the 2015 rank, and returning a value. These values were then clustered into ranges in order to make the graph easier to interpret. Thus, a -10 value means the student's achievement percentile rank was between 0-10 points lower in 2015 than in 2014. While every decile and SGP is displayed on the graph, more than one student can be represented by each point plotted.

Figure 1. Distribution of percentile rank changes, in deciles, by SGP.



Bivariate analysis between SGP and change in percentile rank showed a strong, significant positive correlation ($r = .851$, $p < .01$), indicating that as the SGP increases, so does the change in percentile rank. However, figure 1 tells a more complicated story about individual students, as it uncovers a more nuanced relationship between calculated SGP and achievement (percentile rank). For example, the student labelled with the number 1 in the graph performed at the same level (2) in consecutive years, but exhibited an improvement in percentile rank from 40 to 54, and answered more questions correctly (41 versus 35).

However, despite this student's improvement in standing relative to his peers, his SGP is only 37. Incidentally, this is a dangerously low contribution toward a teacher's mean growth percentile used to determine effectiveness. By contrast, the student labelled with the number 2 has an SGP of 62, yet shows a downward achievement trend as measured by raw scores, scale score and percentile rank. However, because the

MGP is above 50, the measure suggests this student exhibited substantial growth (attributed to the teacher).

Table 2 also illustrates data from the 16-district data set. The four students represented in this table have consistently achieved at the highest performance level on the math tests (level 4). However, their most recent SGPs range from a low of 11 to a high of 95. The gray boxes highlight the "outlier" test scaled scores that drive the 2015 SGP. The first student out-performed his "typical" scoring pattern in 2014 (note 2012 reflects an older, pre-Common Core test scale), resulting in higher-than-typical predicted score of 373.4. The student was unable to reach that prediction, and the low SGP reveals that fact. The second student, row 2, shows a high degree of consistency. This student's scoring pattern falls comfortably along the line predicted by the growth model, and the student has an SGP right smack in the middle- 50. This student's test taking pattern is as predicted. The third row shows a student who exceeded prediction slightly, and the fourth row is the converse of the first row—this student's "good" year is in 2015.

Table 2

Examples of Prior-year Tests Influencing Predictions and Growth Scores

	2015 SGP	2015	2014	2013	2013	Predicted Score
Grade 6 Math	11	350	377	347	725	373.4
Grade 8 Math	50	357	349	345	726	357.2
Grade 6 Math	70	376	360	354	725	365
Grade 7 Math	95	374	341	353	742	344.1

Continued examination reveals a pattern that shows what it takes to get low, close-to-mean, or high SGPs. When a student substantially exceeds a predicted score in a single year, and then performs closer to the longer-term average in the subsequent year, his/her SGP reflects a big drop, resulting in a low SGP. This suggests the phenomenon of *regression to the mean* (Healy & Goldstein, 1978). In other words, repeated measures of SGP over time for an individual student with one year of an outlier score will experience an SGP closer to the mean of 50 over the course of multiple testing experiences. Meanwhile, there is a "good year" to be this student's teacher, and a "bad year" (like 2015).

While Table 2 focused on high-performing test takers, Table 3 illustrates two lower performing students. Both exhibit above-average SGPs for the 2015 school year. However, when you examine the test score history, the grade 8 student is persistently low-performing, while the grade 7 student appears to be on a downward trajectory. Common sense would suggest these students are not heading in the right direction, but their SGPs suggest they are.

Table 3

Examples of Strong-SGP Students with Low Performance Levels

	2015 SGP	2015	2014	2013	2013	Predicted Score
Grade 8 Math	73	263	261	231	658	250.5
		Level 1	Level 1	Level 1	Level 2	
Grade 7 Math	53	290	284	294	703	288.8
		Level 1	Level 2	Level 2	Level 3	

Mean SGP Variability

Moving to the district-level data set, descriptive statistics are again presented (Table 4, below). The population exactly matches state-wide averages for free or reduced lunch and special education, and exceeds the state average for ELLs. The overall mean SGP is 46.5, lower than in the larger dataset, and mean SGPs range from 30.5 in grade 8 to 59.3 in grade 5.

.

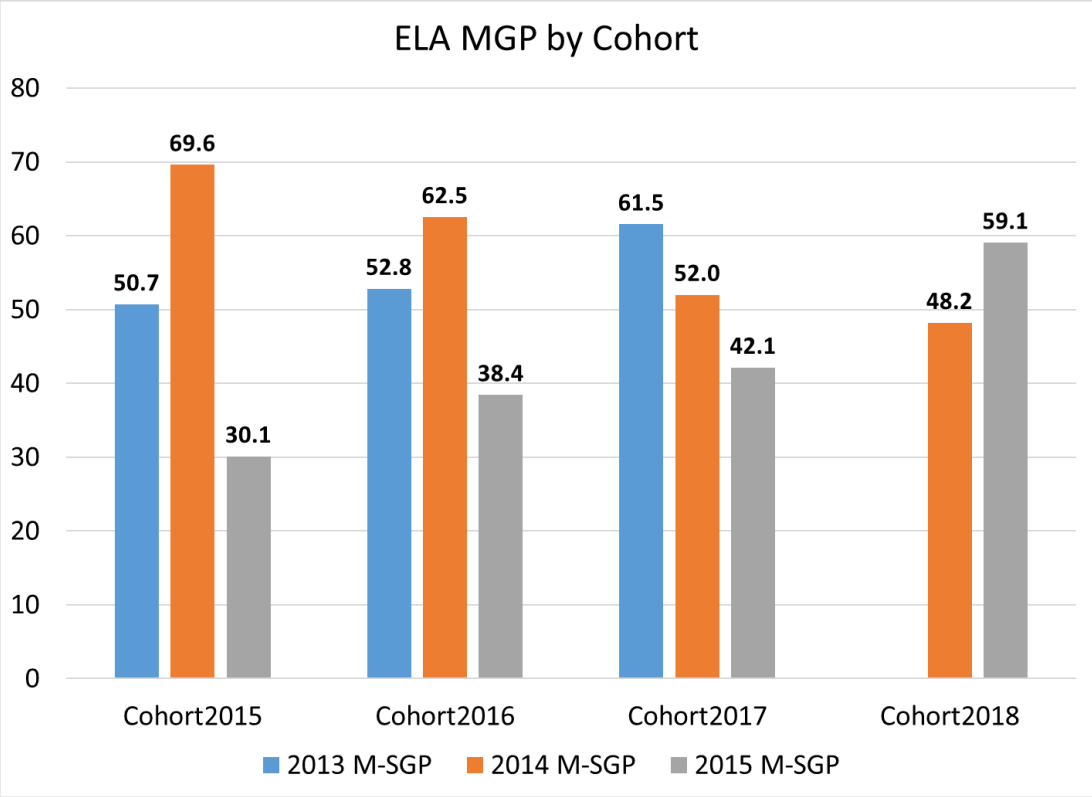
Overall, grades 6, 7 and 8 have decreasing SGPs. This is illustrated more clearly in Figure 2, below. Organized by cohort (i.e., Cohort 2015 are 9th graders in 2015), there are substantial fluctuations in the mean of the SGPs for this cohort over the three years represented. For example, cohort 2015 shows an SGP increase of nearly 20 percentile points, followed by a nearly 40 percentile point drop.

Table 4

Descriptive Statistics for 2015 District-Level Student ELA Data

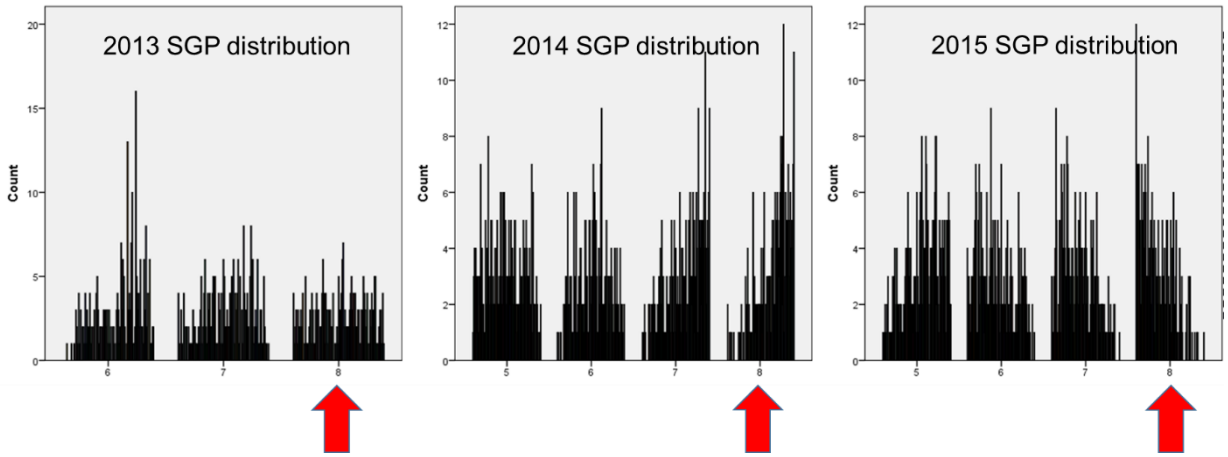
Grade	N	Free or Reduced Lunch		English Language Learners		Students with Disabilities		Mean Student Growth Percentile	
4	288	79		27		28		58.2	
Female	136		40		12		5		61.8
Male	152		39		15		23		55.1
5	312	80		27		35		59.3	
Female	145		35		8		11		61.2
Male	167		45		19		24		57.5
6	266	54		13		19		42.4	
Female	138		27		7		6		43.3
Male	128		27		6		13		41.5
7	277	66		15		21		38.9	
Female	126		33		7		5		41.0
Male	151		33		8		16		37.2
8	265	52		2		15		30.5	
Female	134		24		1		3		32.3
Male	131		28		1		12		28.6
Total	1408	331		84		118		46.5	
		24%		6%		8%			

Figure 2. ELA Mean Growth Percentiles by Cohort, 2013-2015.



The graphs in Figure 3 illustrate in greater detail the distribution of growth scores for the same cohorts depicted in the previous figure. Cohort 2015 is identified by the red arrows, and the left-most graph shows the SGP distribution for 2013, the middle for 2014, and right-most graph 2015. The bars represent the number of students at each SGP value received. While there is a roughly-normal distribution in 2013, this is extremely skewed toward higher SGPs in 2014, and shifts even more drastically in 2015 to mostly lower growth scores.

Figure 3. ELA student growth percentile frequency distributions by year and cohort.



SGPs by Performance Level

Finally, Table 5 (below) reports an analysis of the mean SGP for all students scoring at performance level 1, 2, 3 or 4 in 2015 (16 districts) for ELA and math. There is a nearly 30-point difference in SGP from level 1 performers to level 4 performers in both subjects. All things being equal, the growth model should produce a normal distribution of growth scores across similar student groups, but when results are translated into performance level ranges, it appears that lower performers are systematically receiving lower growth scores than higher performers. Table 6 (below) reports the results of a one-way ANOVA with Bonferroni post-hoc analysis. We see that the large differences between these means is statistically significant (p = .05), meaning there is at least a 95% likelihood that these differences are due to something other than chance alone.

Table 5.

Mean SGP by Performance Level for 2015 ELA and Math

Performance Level	N	Mean ELA SGP	N	Mean Math SGP
1	605	36.1	520	32.6
2	1412	43.8	1001	43.1
3	1509	51.8	1378	49.4
4	841	65.2	1250	60.0
Total	4367	49.6	4149	49.0

Table 6.

SGP One-Way ANOVA results with Bonferroni Post-Hoc Analysis, 2015 ELA and Math Performance Levels

2015 ELA Performance Level		Mean Difference	Std. Error	Sig.	2015 Math Performance Level		Mean Difference	Std. Error	Sig.
1	2	-7.7320*	1.1946	.000	1	2	-10.5037*	1.3651	.000
	3	-15.6685*	1.1830	.000		3	-16.7921*	1.2997	.000
	4	-29.0767*	1.3106	.000		4	-27.4334*	1.3178	.000
2	3	-7.9365*	.9103	.000	2	3	-6.2883*	1.0488	.000
	4	-21.3447*	1.0708	.000		4	-16.9297*	1.0711	.000
3	4	-13.4082*	1.0579	.000	3	4	-10.6414*	.9864	.000

**The mean difference is significant at the 0.05 level.*

Discussion

The main purpose of this study was to address the question, *To what extent does the New York Growth Model for Educator Evaluation provide meaningful student level growth data to inform educator practice and effectiveness*? Analysis of the two data sets, both containing student-level data, raises questions about the meaning of individual SGPs and their potential to influence MGPs (and therefore teacher growth scores) in a manner that can be discordant with evidence of achievement. The following observations based on the above analysis summarize these concerns:

- The relationship between SGP and achievement as measured by percentile rank exhibits a strong positive correlation, but large numbers of individuals exhibit information that can be viewed as contradictory to teachers trying to use this information to determine whether a student has had indeed made meaningful learning gains over the course of the year (figure 1). As the Lederman case has demonstrated, even one missed target (reasonable or not) can negatively influence a teacher rating.
- Year-to-year fluctuations with individual SGPs exhibit regression to the mean over time. This effect is especially evident when students substantially exceed or fail to meet predictions in a given year (tables 2 & 3). Students who far exceed a prediction receive a high SGP in that year, but are likely destined for an equally low SGP in the subsequent year. Non-random assignment of students to teachers can therefore pose a potential threat to the SPGS of teachers who get a disproportionate number of students receiving high SGPs in a given year.
- Regression to the mean also has the potential to occur for entire cohorts of

students in a school (figures 2 & 3). In the single district dataset, large swings in mean SGP resulted in high, then low, teacher evaluation scores in ELA. When a large group of students beats their respective predicted scores, low teacher and principal ratings and scores are likely to follow in the subsequent year.

- Statistically significant differences exist between student growth scores at each of the four student performance levels reported (tables 5 & 6). These differences are substantial, and would cause any reasonable person to recognize the disincentive this could create against wanting to teach a class of low performers.

Policy Implications

New York State's teacher and principal evaluation law, as written, explicitly and implicitly articulates a theory of action that, arguably, communicates the following set of beliefs:

1. Changes in student achievement from one year to the next are an indication of teacher and principal effectiveness.
2. Teacher and principal effectiveness can be differentiated through an analysis of observed student growth on state assessments.
3. Observed differences on these measures allows for identification of bad teachers and principals.
4. Bad teachers and principals will be motivated by their ratings to improve, or to get out of the profession.
5. Better teaching and leadership, or new and better teachers and principals, resulting from this policy will improve student achievement.

The problems outlined in this study paint a confused picture of SGPs as derived from New York's Growth Model for Educator Evaluation. The instability of SGPs experienced by both individual and cohorts of students, coupled with large differences by performance level, raise serious doubts about the ability of this particular model to aid in accomplishing any of the steps in the theory chain above. This study provides enough evidence to warrant a fuller exploration of the model to include an analysis of state-wide SGP trends and patterns, and their implications for the stability of corresponding state-provided growth scores (SPGS).

In the meantime, the State Education Department should consider a moratorium on the use of this model until such a time as a more complete analysis can be done, inclusive of multiple years of SGP data from all districts in the state.

It is particularly important that this occur prior to widespread implementation of the most recent educator evaluation law, which promises to increase the influence of this portion of the evaluation system from 20% to nearly 50% of the overall score. Furthermore, serious effort should be made toward helping teachers and principals make meaning of the confusing, often contradictory measures of student learning based on the state testing program, including SGPs, percentile rankings, scale scores and performance levels. Until this happens, the link between teacher practice and

student achievement on state tests will remain obscured by the confusing output of the growth model.

Finally, it would be prudent for all states and districts using VAMs around the country to carefully examine their use in light of these results. Education leaders and policy makers should establish a mechanism to gauge the degree to which their respective VAMs are meeting the intended policy objectives through empirical studies.

Author Biography

Drew Patrick is a doctoral candidate in the educational program at Manhattanville College. He serves as assistant superintendent for curriculum and instruction in the Bedford Central School District, Westchester County, NY. E-mail: apatrick1258@bcsdny.org

References

Baker, B. D., Oluwole, J., & Green, P. C. (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the race-to-the-top era. *Education Evaluation and Policy Analysis Archives, 21*, 1-71.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Shepard, L. A. (2010). Problems with the use of student test scores to evaluate teachers. EPI briefing paper# 278. *Economic Policy Institute,*

Ballou, D., & Springer, M. G. (2015). Using student test scores to measure teacher performance: Some problems in the design and implementation of evaluation systems. *Educational Researcher, 44*(2), 77-86. doi:10.3102/0013189X15574904

Betebenner, D. W. (2011). A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth projections/trajectories. *The National Center for the Improvement of Educational Assessment. Retrieved from Http://Www.Gadoe.Org/Curriculum-Instruction-and-Assessment/Assessment/Documents/Sgp_technical_overview.Pdf,* 439-450.

Braun, H. (2015). The value in value added depends on the ecology. *Educational Researcher, 44*(2), 127-131. doi:10.3102/0013189X15576341

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014a). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review, 104*(9), 2593-2632.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014b). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review, 104*(9), 2633-79.

Darling-Hammond, L. (2015). Can value added add value to teacher evaluation? *Educational Researcher, 44*(2), 132-137. doi:10.3102/0013189X15575346

Goldhaber, D. (2015). Exploring the potential of value-added performance measures to affect the quality of the teacher workforce. *Educational Researcher, 44*(2), 87-95. doi:10.3102/0013189X15574905

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review, 100*(2), 267-271.

Healy, M., & Goldstein, H. (1978). Regression to the mean. *Annals of Human Biology, 5*(3), 277-280.

Hoffer, T. B., Hedberg, E. C., Brown, K. L., Halverson, M. L., Reid-Brossard, P., Ho, A. D., & Furgol, K. (2011). Final report on the evaluation of the growth model pilot project. *US Department of Education.* Retrieved from https://www2.ed.gov/rschstat/eval/disadv/growth-model-pilot/gmpp-final.pdf

Holloway-Libell, J., & Amrein-Beardsley, A. (2015). "Truths" devoid of empirical proof: Underlying assumptions surrounding value-added models in teacher evaluation. *Teachers College Record, June 29*

McCaffrey, D. F., Lockwood, J., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability. monograph.* ERIC.

New York State Education Department (2012). Guidance on New York State's annual professional performance review for teachers and principals to implement Education Law §3012-c and the Commissioner's regulations. Retrieved from http://www.engageny.org/resource/guidance-on-new-york-s-annual-professional-performance-review-law-and-regulations

New York State Education Department (2015a). New York State Public School Enrollment (2014-15). Retrieved from http://data.nysed.gov/enrollment.php?state=yes&year=2015&grades%5B%5D=03&grades%5B%5D=04&grades%5B%5D=05&grades%5B%5D=06&grades%5B%5D=07&grades%5B%5D=08

New York State Education Department (2015b). Amendment of Subpart 30-2 and Addition of a New Subpart 30-3 to the Rules of the Board of Regents and Section 100.2(o) of the Commissioner's Regulations, Relating to Annual Professional Performance Reviews of Classroom Teachers and Building Principals to Implement Subparts D and E of Part EE of Chapter 56 of the Laws of 2015. Retrieved from https://www.regents.nysed.gov/common/regents/files/meetings/Sep%202015/915p12hea1revised.pdf

Rockoff, J. E., & Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness. *The American Economic Review, 100*(2), 261-266. doi:http://dx.doi.org.librda.mville.edu:2048/10.1257/aer.100.2.261

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics, 125*(1), 175-214. doi:10.1162/qjec.2010.125.1.175

Strong, M., Gargani, J., & Hacifazlioğlu, Ö. (2011). Do we know a successful teacher when we see one? experiments in the identification of effective teachers. *Journal of Teacher Education, 64*(4), 367-382. doi: 10.1177/0022487110390221

Tyler, J. H., Taylor, E. S., Kane, T. J., & Wooten, A. L. (2010). Using student performance data to identify effective classroom practices. *The American Economic Review, 100*(2), 256-260. doi:http://dx.doi.org.librda.mville.edu:2048/10.1257/aer.100.2.256