Is "Effective" the New "Ineffective"?
A Crisis with the New York State Teacher Evaluation System

- by Kenneth Forman, Ph.D., and Craig Markson, Ed.D.

Abstract

The purpose of this study was to examine the relationship among New York State's APPR teacher evaluation system, poverty, attendance rates, per pupil spending, and academic achievement. The data from this study included reports on 110 school districts, over 30,000 educators and over 60,000 students from Nassau and Suffolk counties posted on the New York State Education Department's Data website. The results of this study showed that poverty had a strong negative correlation with performance on the New York State English Language Arts (ELA) and Mathematics assessments among students in grades 3-8. As poverty went up, performance on the State assessments went down. Poverty accounted for over 60 percent of the variance on student performance on both State assessments. The school districts' APPR teacher evaluation ratings had weak to conflicting correlations with student achievement. The school districts' percent of teachers rated "highly effective" had a positive correlation with student achievement. However, the strength of the relationship was weak, accounting for only 12.53 and 10.76 percent of the variance on student success on the English Language Arts and Mathematics examinations respectively. The school districts' percent of teachers rated "effective" had a negative correlation with student achievement. As the percent of teachers rated "effective" went up, student performance on the State assessments went down. The implications of this study suggested that legislators, State education departments, and school districts would better serve students by allocating recourses toward programs that alleviate the detrimental effects that poverty has on academic achievement.

I. Purpose

During the 2011-2012 public school year, New York State implemented a revised teacher evaluation system, the Annual Professional Performance Review (APPR). As was the case with other States' teacher evaluation systems, the APPR has been controversial throughout its implementation (National Center for Education Evaluation and Regional Assistance, 2014; New York State Education Department, 2011). Proponents and critics debated the impact the APPR would have on student achievement (Futscher, 2014; Leonardatos, & Zahedi, 2014). Prior studies suggested that other factors such as poverty, attendance rates, and per pupil spending were more important determinants of student achievement (Arthurs, Patterson, & Bentley, 2014; Hermes, 2005; Jefferson, 2005). As a result, the purpose of this study was to examine the relationship among New York State's APPR teacher evaluation system, poverty, attendance rates, per pupil spending, and academic achievement.

II. Theoretical Framework

Annual Professional Performance Review

On May 28, 2010, New York Governor David Paterson signed Chapter 103 of the Laws of 2010, which added section 3012-c to the Education Law, establishing a comprehensive evaluation system for teachers, requiring classroom teachers to receive an annual professional performance review rating (APPR) from a composite effectiveness score with a score of "highly effective," "effective," "developing," or "ineffective." The composite score was to be determined as follows: (a) 20% based on student growth on State assessments or other comparable measures of student growth (increased to 25% upon implementation of a value-added growth model), (b) 20% based on locally-selected measures (SLOs - student learning objectives or MOSL- measures of student learning) that were rigorous and comparable across classrooms (decreased to 15% upon implementation of value-added growth model) and (c) 60% based on other measures of teacher effectiveness, reflecting observation of teacher performance using a State approved evaluation rubric. For the 2011-2012 school year, the law only applied to classroom teachers of the common branch subjects, English Language Arts or Mathematics in grades 3-8. In the 2012-2013 and 2013-2014 school years, the law applied to all classroom teachers and building principals. The APPR was designed to be a significant factor in employment decisions such as promotion, retention, tenure determinations, termination, and supplemental compensation, as well as a significant factor in teacher professional development. Scoring ranges that determined teachers' performance levels were developed as a result of negotiations between school district and union (NYSED, 2014). Early in 2015, the New York State Legislature passed a law altering the APPR requirement so that student

performance still plays a role in teacher rating. This new law prescribes how teachers might be rated using a matrix (NYSED, 2015). The New York State Board of Regents, the State education governing body, has the charge of defining critical elements for implementation.

Teacher Evaluation and Student Achievement

There are a variety of concerns with using student achievement data on both State and local assessments to evaluate teachers. One of the main problems in tying test scores to teacher evaluation is determining if some teachers are simply more effective at helping students achieve, or if some teachers happen to have more able students in their classroom. Darling-Hammond, Amrein-Beardsley, Haertel, and Rothstein (2012) found that student achievement could be influenced by much more than simply a teacher's effectiveness. Class size, curriculum materials available, availability of learning materials and technology resources, and staffing of specialists in a school building can all affect student achievement. Concomitantly, challenges in student home life, family income, and issues in a community can likewise affect student achievement, as well as individual student needs, attendance, student health, and culture. A student's prior teacher and schooling, differential summer learning loss and assessment type were also factors that can affect student achievement that may be outside of the teacher's control (Darling-Hammond, et al. 2012). In a separate study, Darling-Hammond (2015) reported that teachers became more effective as they received feedback from standards-based observations and as they developed ways to evaluate their students' learning in relation to their practice.

However, there seem to be inaccuracies and potential validity issues with using value-added data regarding how much the value-added portions of composite teacher evaluations should be weighted. Although many States are implementing value-added teacher evaluation systems, there have been alignment concerns between what current research deems best practice and what has been pushed onto many schools because of initiatives that demand more accountability with teacher evaluations (Snyder et al., 2012).

Teacher effectiveness has been linked to instruction by combining them into a single index to balance out the effect of differences in student background. However, there has been little empirical evidence to indicate how this combined index might weight each measure toward a composite teacher evaluation. According to the *Measures of Effective Teaching* (MET), a balanced approach was most sensible when assigning weights to form a composite teaching measure, as too much emphasis on any one piece of a teacher's composite score could be misleading (Bill and Melinda Gates Foundation, 2013). A teacher's composite score was comprised of student achievement gains on State tests, student survey responses and observations using Charlotte Danielson's Framework for Teaching rubric (Danielson, 2007). The MET study

correlated these factors with student achievement; for example, the 2009-2010 composite measure of teaching accurately predicted the 2010-2011 student performance. Additionally, students who were randomly assigned to a teacher previously rated "effective" performed better on State assessments than expected that year based on individual students' past exam scores. On the other hand, students who were randomly assigned to a teacher that was identified as "less effective" actually achieved a lower grade than predicted based on their own individual past exam scores. Concomitantly, the MET researchers reported that there were a variety of challenges in using test scores to evaluate teachers (Bill and Melinda Gates Foundation, 2013).

Another study in a large western school district analyzed teacher evaluation scores based on Danielson's Framework for Teaching by comparing student achievement measures. Analysis involved reviewing teacher evaluation scores based on an observation rubric with district and State examinations in reading, mathematics, and a composite test on reading and mathematics. This study provided some evidence of a positive relationship between teacher performance, as measured by the evaluation system, and student achievement (Kimball et al., 2004).

Milanowski (2004) conducted a similar study around the same time, analyzing the relationship between teacher evaluation scores and student achievement on district and State examinations in reading, mathematics, and science in another large mid-western school district. The results of this study indicated that scores from a rigorous teacher evaluation system using a value-added framework could be significantly related to student achievement.

Berliner (2013) reported that there were many intrinsic problems with value-added evaluation of teachers, especially issues with the testing process itself. In his discussion on the lack of instructional sensitivity of test items, he reaffirmed that higher social class students had higher passing rates per item and lower social class students had lower passing rates per item, independent of the teacher's ability to teach (Berliner, 2013).

Haertel (2013) explained that no statistical manipulation was able to assure fair comparisons of teachers working in very different schools, with very different students under very different conditions. However, the MET study indicated that teachers had a major influence on student learning, especially when multiple measures helped identify how a teacher contributed to student learning. When teacher actions were unstable, teacher value-added scores were unstable. The researchers found that teacher behavior in classrooms varied because of a variety of factors, including: constantly changing student behavior, the need to teach multiple school subjects each day, daily changes in scheduling, and daily differences in absenteeism by students, teachers, aides and support personnel. The MET study also indicated that composite

evaluations that combined different aspects of teacher evaluation were better than using just one, teacher observers needed rigorous training and teachers should be observed multiple times per year by multiple observers. Additionally, the MET study supported that student gains needed to be adjusted to account for differences in the students. When the researchers found a correlation of student achievement with teacher ratings, that correlation was weak and quite low (Bill and Melinda Gates Foundation, 2013).

Marshall (2013) identified six factors that he felt did not support the relationship of teacher ratings with student achievement and standardized testing. He suggested that standardized tests were never designed to evaluate teachers. Moreover, districts would need to collect three years of value-added scores to reduce "noise" from the data and fear of negative consequences could lead to teachers spending an inordinate amount of time on test prep. Additionally, evaluating teachers on the basis of test results could have a negative effect on collegiality. Finally, he indicated that standardized test data were only available for 20% of teachers and praising or critiquing teachers failed to take into account work done by "pullout" teachers, specialists, tutors, or previous grades. Marshall emphatically concluded it was problematic to use standardized test scores to evaluate teachers (Marshall, 2013).

Poverty and Attendance

Studies by Darling-Hammond et al. (2012) and Darling-Hammond (2015) revealed that students' achievement and measured gains were influenced by much more than any individual teacher. A multitude of factors were identified and included the effects of poverty, such as: home and community supports or challenges, individual student needs and abilities, health and attendance, peer culture and achievement, differential summer learning loss which especially affected low-income children, and the specific tests used which emphasized some kinds of learning and not others, and which rarely measured achievement that was well above or below grade level (Darling-Hammond, 2015; Darling-Hammond, et al., 2012).

Hershberg et al., (2004) indicated that it was impossible to fully separate out the influences of students' other teachers as well as school conditions on students' reported learning. No single teacher accounted for all of a student's learning. Prior teachers had lasting effects both positive and negative on students' later learning. By following individual students over time, value-added assessment was influenced by student background characteristics over which schools had no control and that tended to bias test results (Hershberg et al., 2004).

Linda Darling-Hammond (2015) reported that the US educational system was one of the most segregated and unequal in the industrial world because of our high rates of childhood poverty and homelessness and food insecurity that were not randomly distributed across communities. Moreover, schools and districts have unequal funding so that teachers working in lower income communities often have fewer resources to serve concentrations of students with greater need (Darling-Hammond, 2015).

In a study of value-added teacher effectiveness by Newton et al., (2010), the researchers found that even though three of the five models controlled for student demographics as well as students' prior test scores, teachers' rankings were nonetheless significantly and negatively correlated with the proportions of students they had who were English language learners, free lunch recipients, or Hispanic, and were positively correlated with the proportions of students they had who were Asian or whose parents were more highly educated. The researchers' findings highlighted the challenge inherent in developing a value-added model that adequately captured teacher effectiveness when teacher effectiveness itself was a variable with high levels of instability across contexts (i.e., types of courses, types of students, and year). Even in models that controlled for student demographics as well as students' prior test scores, teachers' rankings were nonetheless negatively correlated with the proportions of students they had who were English language learners, free lunch recipients or Hispanic. Rankings were positively correlated with proportions of students who were Asian or whose parents were more highly educated. The default assumption in the value-added literature was that teacher effects were a fixed construct that was independent of the context of teaching (e.g., types of courses or student demographic compositions in a class) and stable across time. The researchers found that empirical exploration of teacher effectiveness rankings across different courses and years suggested that this assumption was not consistent with reality. Correlations indicated that even in the most complex models a substantial portion of the variance in teacher rankings was attributable to selected student characteristics (Newton et al., 2010).

Per Pupil Spending

The New York State Department of Finance conducted a study toward better understanding of the relationships among instructional expenditures per pupil, district need, and educational performance. This study examined expenditures, district need and academic performance from different perspectives to develop some insights and a better understanding of these relationships. The department concluded: (a) adjusting expenditures per pupil for need and cost was a productive approach to understanding the relationships among expenditures, student need and academic performance; (b) after accounting for cost and need, expenditures per pupil can make a difference in academic performance; and (c) perhaps the greatest challenge was to improve educational effectiveness in high needs districts. High needs districts need to increase instructional expenditures on a per pupil basis to improve academic performance (NYS Department of Finance, 2004).

III. Data Sources

The data from this study were obtained from the New York State Education Department Data Site (2015) for the 2013 to 2014 school year. State reporting on 110 school districts from Nassau and Suffolk Counties, New York were included in this study. There were 15 school districts located in Nassau and Suffolk Counties that were excluded from this study for having a population of less than 50 teachers. The New York State Education Department Data Site (2015) was the source of the following data: (a) the number and percent of students collecting free and reduced lunch; (b) the percent of average daily student attendance; (c) the numbers of educators and their APPR teacher rating percentages; and (d) grades 3-8 student achievement as indicated by levels 3 and 4 on State English Language Arts and Mathematics examinations. The source of data to determine per pupil spending was the tax levy portion of the 2014 school district budgets obtained from the *Newsday* website ("Long Island school districts' tax plan," n.d.).

IV. Method

Student achievement was measured by performance on standardized State examinations in English Language Arts and Mathematics, grades 3-8. There were 4 reporting levels. Level 1 was considered exceedingly below grade level expectations. Level 2 was

considered students' performance approaching grade level, and Levels 3 to 4 were students performing on grade level and above. Student achievement was the dependent variable and measured by the percent of students obtaining Levels 3 to 4 on the English Language Arts and Mathematics State examinations. Poverty was identified as the percent of students receiving free and reduced lunch district-wide. Attendance was indicated as the percent of average daily attendance for the entire school district. Teacher Performance included the percent of teachers rated on each category of the district's Annual Personnel Performance Review. The Annual Personnel Performance Rating (APPR) evaluation system categorizes teacher effectiveness according to four performance levels: Level 1 - "ineffective," Level 2 - "developing," Level 3 - "effective" and Level 4 - "highly effective." Per pupil spending was determined by dividing by the tax levy school district budget by pupil population. The tax levy was the amount of funding available to districts through direct taxation of its residents, not influenced by a variety of other funding sources, and thus provided a clear per pupil spending amount.

Two correlation analyses were conducted to determine if school districts' free and reduced lunch (poverty) , attendance rate, teacher rating -"highly effective," "effective," "developing," and "ineffective" APPR percentages, and per pupil spending were related to the percent of its students scoring Level 3 and/or 4 on the English Language

Table 1
Correlations with ELA Level 3 or 4 Achievement Percentage (N = 11 - 110)

		ELA Level 3 or 4	Free & Reduced Lunch	Attendance Rate	Highly Effective APPR	Effective APPR	Developing APPR	Ineffective APPR
Free & Reduced Lunch	r	-0.777**						
	r^2	60.37%						
Attendance Rate	r	0.469**	-0.456**					
	r^2	22.00%	20.79%					
Highly Effective APPR	r	0.354**	-0.265**	0.158				
	r^2	12.53%	7.02%	2.50%				
Effective APPR	r	-0.331**	0.241*	-0.142	-0.987**			
	r^2	10.96%	5.81%	2.02%	97.42%			
Developing APPR	r	-0.113	0.127	-0.079	-0.421**	0.2		
	r^2	1.28%	1.61%	0.62%	17.72%	4.00%		
Ineffective APPR	r	-0.667*	0.679*	-0.651*	-0.188	-0.169	0.288	
	r^2	44.49%	46.10%	42.38%	3.53%	2.86%	8.29%	
Per Pupil Spending	r	0.562**	-0.4**	0.21*	0.134	-0.119	-0.009	-0.329
	r^2	31.58%	16.00%	4.41%	1.80%	1.42%	0.01%	10.82%

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

Arts and Mathematics examinations. A Pearson Product-Moment correlation analysis, with a two-tailed test of significance with alpha set at .05, was used to analyze the relationships between the variables.

V. Results

Table 1 illustrates the results for the correlations with ELA Level 3 or 4 achievement.

The percent of students receiving free and reduced lunch had a statistically significant relationship with the percent of students achieving Level 3 or 4 on the ELA assessments. There was an inverse relationship, accounting for 60.37 percent of the variance: as the percent of students receiving free and reduced lunch increased, the percent of students achieving Level 3 or 4 achievement substantially decreased. The attendance rate also had a statistically significant relationship with the ELA assessment rate. Here, there was a positive correlation, accounting for 22 percent of the variance on the percent of students receiving Level 3 or 4 on the ELA assessments. The percent of teachers rated "highly effective" had a statistically significant and positive correlation with student ELA scores, accounting for 12.53 percent of the variance. The "effective" teacher rating also had a statistically significant but negative correlation with the ELA assessments, accounting for 10.96 percent of the variance.

The "developing" APPR rating did not have a statistically significant relationship with ELA assessments, p > .05. The "ineffective" APPR rating had a statistically significant and negative correlation with the ELA assessments, accounting for 44.49 percent of the variance. Finally, the districts' per pupil spending had a statistically significant and positive correlation, accounting for 31.58 percent of the variance.

Table 2 displays the results for the correlations with Level 3 or 4 achievement on the Mathematics assessments. The percent of students receiving free and reduced lunch had a statistically significant relationship with the percent of students achieving Level 3 or 4 on the New York State Mathematics assessments. There was an inverse relationship, accounting for 62.57 percent of the variance: as the percent of students receiving free and reduced lunch increased, the percent of students achieving Level 3 or 4 achievement substantially decreased. The attendance rate also had a statistically significant relationship with the Mathematics assessment rate. Here, there was a positive correlation, accounting for 23.91 percent of the variance on the percent of students receiving Level 3 or 4 on the Mathematics assessments. The percent of teachers rated "highly effective" had a statistically significant and positive correlation with student Mathematics scores, accounting for 10.76 percent of the variance. The "effective" teacher rating also had a statistically significant but negative correlation with the Mathematics assessments, accounting

Table 2

Correlations with Mathematics Level 3 or 4 Achievement Percentage (N = 11 - 110)

		Math Level 3 or 4	Free & Reduced Lunch	Attendance Rate	Highly Effective APPR	Effective APPR	Developing APPR	Ineffective APPR
Free & Reduced Lunch	r	-0.791**						
	r^2	62.57%						
Attendance Rate	r	0.489**	-0.456**					
	r^2	23.91%	20.79%					
Highly Effective APPR	r	0.328**	-0.265**	0.158				
	r^2	10.76%	7.02%	2.50%				
Effective APPR	r	-0.295**	0.241*	-0.142	-0.987**			
	r^2	8.70%	5.81%	2.02%	97.42%			
Developing APPR	r	-0.202	0.127	-0.079	-0.421**	0.2		
	r^2	4.08%	1.61%	0.62%	17.72%	4.00%		
Ineffective APPR	r	-0.634*	0.679*	-0.651*	-0.188	-0.169	0.288	
	r^2	40.20%	46.10%	42.38%	3.53%	2.86%	8.29%	
Per Pupil Spending	r	0.496**	-0.4**	0.21*	0.134	-0.119	-0.009	-0.329
	r^2	24.60%	16.00%	4.41%	1.80%	1.42%	0.01%	10.82%

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

for 8.76 percent of the variance. The "developing" APPR rating did not have a statistically significant relationship with Mathematics assessments, $p > .05$. The "ineffective" APPR rating had a statistically significant and negative correlation with the Mathematics assessments, accounting for 40.2 percent of the variance. Finally, the districts' per pupil spending had a statistically significant and positive correlation, accounting for 24.6 percent of the variance.

VI. Conclusions

The Annual Professional Performance Review (APPR) rating that had the strongest correlation with student success, Levels 3 or 4, on both the English Language Arts and Mathematics examinations was the "ineffective" category, accounting for 44.49 and 40.2 percent of the variance on the assessments respectively. Predictably, this had a negative correlation with students' performance on both State assessments. However, only 11 of the 110 districts included in this study had reporting in the "ineffective" category. The other 99 districts had zero percentage reporting. While the "highly effective" category had all 110 districts reporting various percentages of its teachers in this category, it only accounted for 12.53 and 10.76 percent of the variance on student success on the English Language Arts and Mathematics examinations respectively. The "effective" category also had all 110 districts reporting various percentages of its teachers in this category. However, what was surprising was the inverse relationship that the "effective" APPR category had with the student achievement success rates, Level 3 and 4. With only 11 school districts reporting "ineffective" and the inverse relationship that "effective" had with student achievement, "effective" has become the new "ineffective." This was probably caused from the underreporting of "ineffective" and "developing" categories, which had only 53 school districts reporting some percentage of its teachers in these categories and as such, the results were skewed.

The real crisis with the New York State teacher evaluation system was that it overshadowed the most important problem of poverty and its harmful effects on student achievement. The percent of students receiving free and reduced lunch, which was used to measure poverty, accounted for a whopping 60.37 percent of the variance on student success on the English Language Arts examinations and 62.57 percent on the Mathematics examinations. The correlation analyses also revealed that as poverty went up, attendance rates went down. Lower school attendance also put downward pressure on student success on the State assessments. There was a positive correlation of student attendance and student achievement. The results of this study showed districts that had a high average daily attendance also evidenced higher levels of student achievement. Finally, there was a positive correlation of student achievement with per pupil spending. The higher the per pupil spending by district, the greater the student achievement.

VII. Implications of the Research

As a result of Race to the Top federal funding, New York State (along with other RTT award recipient States) adopted a paradigm for teacher evaluation involving multiple measures for determining teacher effectiveness. Likewise, with the pending renewal of the federal Elementary and Secondary Education Act (ESEA) policy makers will face an important question: Can teacher effectiveness be reliably measured using value-added metrics to evaluate teachers and hold them accountable? This dilemma is not easily resolved, but after looking at the data from 110 school districts across Long Island with over 67,000 students and 32,000 teachers there are some obvious suggestions.

Use enhanced teacher observation protocols with multiple trained evaluators and downplay the importance of testing for teacher evaluation since value-added metrics have proven to be unreliable and an inaccurate predictor of teacher performance. Rather than relying on these metrics for determining growth in student achievement, other evidence should be considered. Perhaps using formative English Language Arts or Mathematics assessments or looking at growth in students' written work according to a defined rubric might have greater value. For English Language Learners, perhaps looking at growth over the school year on vocabulary acquisition might prove more worthy.

As more demands are placed on principals to evaluate their teachers in an objective and standardized format, principals will be forced to lean on their teachers to perform other important duties, such as curriculum and professional development, and to lead work in different structures within the school, such as professional learning communities or instructional rounds. A new breed of teachers will evolve, "teacher leaders" who would assume responsibility as leading learners for their schools, leading their colleagues collaboratively to maximize student achievement.

Moreover, if the results of this study remain consistent with future studies, legislators, State education departments, and school district leaders throughout the country should focus more of their attention on developing programs that alleviate the detrimental effects that poverty has on student achievement. A variable that accounts for over 60 percent of the variance on student achievement should not be ignored.

References

Arthurs, N., Patterson, J., & Bentley, A. (2014). Achievement for students who are persistently absent: Missing school, missing out? *The Urban Review*, 46(5), 860-876.

Berliner, D. C. (2013). Problems with value-added evaluations of teachers? Let me count the ways! *Teacher Educator*, 48(4), 235-243.

Bill and Melinda Gates Foundation. (2013). Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET Project's three-year study. *Policy and Practice Brief*. MET Project.

Danielson, Charlotte. (2007). Enhancing professional practice: A framework for teaching. 2nd Edition. *Association for Supervision and Curriculum Development.*

Darling-Hammond, L. (2015). Can value-added add value to teacher evaluation? *Educational Researcher*, Vol. 44.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating Teacher Evaluation. *Phi Delta Kappan*, 93(6), 8-15.

Nassau County Council of School Superintendents. (2013 - 2014). *Directory of public schools, Nassau County.* Nassau BOCES Publications, Syosset, New York.

Futscher, M. R. (2014). Principals' perceptions about the impact of the New York State Annual Professional Performance Review legislation on the instructional practices of New York public schools. *Dissertation Abstracts International Section A*, 75.

Haertel, E. (2013). Reliability and validity of inferences about teachers based on student test scores. *Educational Testing Service*, Princeton NJ.

Hershberg, T., Simon, V. A., & Lea-Kruger, B. (2004). The revelations of value-added: an assessment model that measures student growth in ways that NCLB fails to do. *School Administrator,* 61(11), 10

Hermes, M. (2005). Complicating discontinuity: What about poverty? *Curriculum Inquiry,* 35(1), 9-26.

Jefferson, A. L. (2005). Student performance: Is more money the answer? *Journal of Education Finance*, 31(2), 111-124.

Kimball, S. M., White, B., Milanowski, A. T., & Borman, G. (2004). Examining the Relationship between Teacher Evaluation and Student Assessment Results in Washoe County. *Peabody Journal of Education,* vol. 79.

Leonardatos, H., & Zahedi, K. (2014). Accountability and "Racing to the Top" in New York State: A Report from the Frontlines. *Teachers College Record*, 116(9), 1-23.

Long Island school districts' tax plan. (n.d.). Retrieved March 9, 2015, from http://data.newsday.com/long-island/data/education/school-taxes-2015/

Marshall, Kim. (2013) *Re-Thinking teacher supervision and evaluation: How to work smart, build collaboration, and close the achievement gap,* 2nd Edition. Jossey-Bass.

Milanowski, Anthony. (2004) The Relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education,* vol. 79.

National Center for Education Evaluation and Regional Assistance. (2014). State requirements for teacher evaluation policies promoted by Race to the Top. NCEE Evaluation Brief. NCEE 2014-4016. *National Center for Education Evaluation and Regional Assistance*.

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. Value-Added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Educational Policy Analysis Archives*, vol. 18 (2010)

New York State Department of Finance. (2004). *Towards an Understanding of the Relationships among Student Need, Expenditures and Academic Performance.*

New York State Education Department. Learning Summit. (2015). Retrieved March 10, 2015, from: http://www.nysed.gov/common/nysed/files/summary-of-2015-evaluation-changes.pdf

New York State Education Department. (2012) *Approved Teacher Practice Rubrics for New York State*. Retrieved March 7, 2015, from:
http://usny.nysed.gov/rttt/teachersleaders/practicerubrics/

New York State Education Department Data Site. (2015). Retrieved March 9, 2015, from http://data.nysed.gov/

New York State Education Department. (2012). *Guidance on New York State's Annual Professional Performance Review for teachers and principals to implement Education law §3102-c and the Commissioner's regulations*. (2012) Retrieved March 11, 2015, from: http://engageny.org/wpcontent/uploads/2012/05/APPR-Field-Guidance.pdf

SCOPE. (2013-2014). *SCOPE Directory of Suffolk County public schools and educational organizations serving Long Island.* Smithtown, New York.

Snyder, J. E., Hadfield, T. E., & Hutchison-Lupardus, T. (2012) Assessing state models of value-added teacher evaluations: alignment of policy, instruments, and literature-based concepts. (Order No. 3516314, Saint Louis University). ProQuest Dissertations and Theses, 168. (2012). Retrieved from: http://search.proquest.com/docview/1029146032?accountid=14172.

Kenneth Forman, Ph.D., is Program Coordinator in the Educational Leadership Program at Stony Brook University, Long Island, NY.

Craig Markson, Ed.D., is a graduate of Dowling College, in Oakdale, NY, and Assistant Dean in the School of Professional Development at Stony Brook University, Long Island, NY.