The Evaluation of Rhode Island Public High School Teachers: The Impact on Students

The Honors Program
Senior Capstone Project
Student’s Name: Stephen Lamontagne
Faculty Sponsor: Alan Olinsky
April 2016

Table of Contents

ABSTRACT

In 2012, the state of Rhode Island began the full implementation of a high-stakes teacher evaluation system. Its purpose is to increase teacher accountability and to improve student performance. However, a significant amount of literature casts doubt about the effectiveness and validity of teacher evaluation. This paper utilizes statistical methods including regression and decision trees in order to determine whether or not there is a relationship between teacher evaluation in Rhode Island and student performance, using RI Department of Education Data for each school from 2008-2015. Furthermore, this presentation investigates other factors that affect schools, to see if changes in student performance can be explained by factors other than the teacher evaluation system, such as discipline, the student-teacher ratio, and student demographics.

INTRODUCTION

Over the last fifteen years, there has been a movement in the fields of primary and secondary education to introduce new and innovative methods to improve student learning and performance in order to ensure a higher quality of education for students. Although public education is very localized across the nation, this national push led to an increased degree of uniformity in American education while still maintaining regional differences. Two very prominent examples of measures to improve student performance are the widespread use of standardized testing in the aftermath of the 2001 No Child Left Behind Act and the Common Core curriculum promoted by the Obama administration. However, a third common method being used to improve student performance is the evaluation of teachers, which has been implemented in schools across the country in a variety of different ways, due to many states and school districts adopting the idea. Although the core idea of evaluating teachers is now common in many places in the United States, the exact methods used vary greatly. Several different methods, including value-added models, subjective evaluation and the use of standardized testing in high-stakes educator evaluation have been employed. While significant research has been undertaken to determine the benefits and drawbacks of these evaluation methods, the state of Rhode Island teacher evaluation in the fall of 2012, using a system that combines several common methods of evaluating teachers in the fall of 2012. This paper evaluates existing literature on different types of teacher evaluation in order to determine how effective the process tends to be, and utilizes statistical methods in order to determine whether or not the specific method of teacher evaluation that has been implemented in Rhode Island has in fact benefited the students in Rhode Island public high schools.

LITERATURE REVIEW

Overview

Throughout the history of public education in the United States, there have constantly been new concepts introduced that have been designed to improve the education of the students. These have varied from methods that dictated what should be taught and how to teach it, to methods that gave teachers more autonomy, and to methods that put an administrator, such as the principal, in a position to supervise what a teacher was doing in the classroom. Tenure and pay scales based on length of service became normal.

However, in the late 1990s, a movement to begin evaluating teachers was formed, so the concept of high-stakes evaluation of teachers is a relatively recent concept. Congress began to pass legislation relating to standards for education in the late 1990s, and in 1999, the National Education Summit began to discuss paying teachers for performance (Holt, 2001, 312). The first significant implementation of a standards-based evaluation system for secondary school educators occurred in 2001, when the landmark No Child Left Behind Act was passed by the Bush administration. The law mandated the provision of equal access to education for all students regardless of background, and strongly pushed the implementation of high-stakes standardized testing as a method to ensure that all students were having their educational needs being met (Caillier, 2010, 58). Although standardized testing of students has existed for a significantly longer period of time than high stakes evaluation of teachers has, it was with the passage of No Child Left Behind that standardized testing began rapidly turning into the high-stakes assessments that they are today.

Federal government education programs provide strong incentives for states and school districts to adopt new strategies, but do not require the states to have uniformity in education. The result of this is that each state has developed its own version of standardized testing in the intervening years. The evaluation methods for teachers were initially developed at the level of the school district, which resulted in an even more varied breakdown of evaluation methods being used to make decisions about teacher quality. Some states saw that No Child Left Behind failed to indicate a way to determine how much an individual teacher enabled a student to progress, so some states began to implement value-added models to remedy this (Caillier, 2010, p. 58). Other states and districts began using other methods, including high-stakes testing and subjective evaluation. Although the Common Core curriculum has recently been adopted, it does not tend to affect the evaluation of the teachers, a trend that is continuing to become more refined as districts modify their existing methods of evaluation. There is only current literature on the topic of teacher evaluation, since it is such a recent development, and major issues such as determining the most effective method of evaluating a teacher is still in progress.

Review

The contemporary literature on public high school education all indicate that the methods that were first designed to encourage student performance may in fact have a negative impact on student education. There is broad discussion of issues in the current system, although there is general dissent on what an optimum solution would be for improving it. The high degree of emphasis given to high-stakes testing results, according to most of the researchers, limits teacher ability to teach a variety of material, but it is important to note that teachers should not simply eliminate this content. Furthermore, the research indicates that many teachers respond to the importance of the high-stakes testing by focusing specifically on student performance on those tests, including drills on sample questions and teaching at the same scope as the test. Gunzenhauser (2003) takes this argument a step further than most and argues that not only does the system of high-stakes testing limit the flexibility of educators but that the measures of standardized tests are fallible and not necessarily the most accurate measures of student success, when a variety of these measures are needed to fully understand education. He makes the point that the scores from one individual test are only an approximation of one type of student success, of which a variety are needed. Furthermore, in a position also held by Lazear (2006), school districts will tend to focus solely on areas that are tested when standardized testing is made high-stakes for students. At the same time, they will avoid teaching other areas of the curriculum, thereby weakening students in some subject areas. He indicates that the best use of these scores are not to be high-stakes measures but rather feedback on curricula, to be used for the purpose of improving the curricula. He therefore implies that rating individual teachers based on high-stakes testing would lead to biased results that might not actually be representative of actual teacher performance.

The literature tends to be mostly written by university professors and administrators, with the notable absence of authors who directly participate in the education of high school students, for the purpose of producing a stronger education system. The majority of articles tend to focus on the accuracy of measures of students success, and on the effectiveness of particular evaluation methods in measuring the performance of individual teachers. There is very little research investigating the impact evaluating teachers by a particular method has on student performance. Researchers tend to break down teacher evaluations into several main categories, which will be discussed individually below.

High-Stakes Testing

Lazear (2006) defines high-stakes standardized testing as a system where "teachers, administrators, or students are punished for failure to pass a particular exam" (p. 1029), and notes that it is currently an essential piece of public education policy, since the implementation of No Child Left Behind in 2001 (p. 1029). He argues that such a system is effective under particular circumstances, while in others, implementing a high-stakes testing system as an important component of evaluation is instead counter-productive. When it is more difficult for a student to learn, Lazear argues, standardized testing incentivizes those students to learn the material that will be on the high-stakes test, an incentive that would otherwise be lacking. However, for students who tend to be high performers, a high-stakes standardized testing system tends to narrow the curriculum, and in fact causes these students to learn less than they would have otherwise (p. 1042).

Brimijoin (2005) agrees that there are flaws with the high-takes testing systems, but has a significantly different perspective on the issue than Lazear. She argues that the system of mandatory high-stakes testing to evaluate students has incentivized teachers to not teach to the best of their ability to cover all students' needs, but rather uniformly cover standardized test content with the class so their students appear to perform strongly, even if they have to teach to the test (p. 255). She promotes the idea of differentiation, or tailoring to each individual student's needs, as the ideal in education, and claims that standardized testing prevents that. No Child Left Behind was designed, in part to increase equal access to education among disadvantaged children, and while Brimijoin asserts that high-stakes testing can in fact do this, she also believes it limits teachers' professional discretion to educate their students in the most effective way possible (p. 256). Ford (2013) found that students from minority groups tend to underperform relative to other students (p. 115), while Brimijoin found that high-stakes testing can make education more equitable across different demographics (p. 256), thus partially addressing the issue realized by Ford. Brimijoin further concludes that simply having high-stakes testing for students is not enough to ensure a high quality education, but further measures should also be implemented in order to ensure that students receive a quality education, and believes that schools should focus on education goals, but not scores on one specific test (p. 257). She does not make any specific recommendations.

Subjective Evaluation

One method that can be used in order to develop some of these alternative goals is to implement some form of subjective method of evaluation for educators, to provide an incentive to build more skills in their students beyond the score on a particular standardized test. Moore and Kuol (2005) look at the benefits that teachers can gain from having both an objective, test-based, method used in their evaluations, when combined with a subjective response from students (p. 69). When the two evaluations align, Moore and Kuol assert, the system functions the same, but can cause additional actions by teachers in response to certain types of feedback. If the objective review is positive, but the subjective one is negative, the teacher may start to try to address small issues that may not really affect the quality of teaching. In the opposite scenario, the teacher gains confidence that they can do some things correctly, and it more accurately identifies areas of improvement for the teacher (p. 69).

Moore and Kuol (2005) also identify the potential issue of teachers who receive positive feedback in any evaluation system feeling confident and secure as a result of the evaluation results, and therefore are not as careful with education quality in the future. This could negatively impact the education of future students, but this is by no means a definite consequence of the implementation of a teacher evaluation system (p. 68). Brimijoin (2005) agrees with them that there should be a subjective component of the evaluation system, but does not suggest that it be provided by students. Although she does not specifically state what a subjective evaluation could be, she does indicate that it should be through the school system as an entity and not from the students (p. 257). Although both of these papers have solid qualitative arguments, the quantitative investigations of the issues are either poor or completely lacking, an issue that Rockoff and Speroni (2010) address, running a quantitative study of students in New York, indicating that high subjective evaluations are associated with teachers whose students learn more in their classes, in the first two years of their careers (p. 264). They are all in agreement about the importance of including subjective evaluation as a measure of the performance of teachers, since it provides greater flexibility to evaluation systems, and can broaden the scope of the evaluation process. The high-stakes test is not the only form of objective evaluation that subjective evaluation of secondary school educators can be combined with, however, and another common model is the Value-Added Model, which has its own set of advantages and disadvantages.

Value-Added Models

A value-added model, according to Hill, Kapitula, and Umland (2011), is a model that measures student gains over the course of a year in order to assess the effectiveness of a certain school or certain teacher (p. 795). These three writers note several issues with the use of value-added models for important decisions such as evaluating the effectiveness of a particular teacher, noting that factors such as family background and the efforts of other teachers can mistakenly get attributed to one particular teacher, can be difficult to measure, and may be biased by the classroom demographics assigned to an individual teacher. The results of Sass, Semykina, and Harris (2014), provide one of the strongest quantitative studies on a teacher evaluation topic. They completed a large study on the effectiveness of value-added models using educational data from the state of Florida, and agree with the issues noted by Hill, Kapitula, and Umland (p. 35). The study showed that there are definite issues with separating out the impact that individual current teachers, as well as prior teachers, have had on students' learning, and it is suggested that value-added models not be used as a sole indicator of the performance of a teacher.

Furthermore, the study indicated that better performance on the value-added model correlated with better results on subjective reviews from school administrators (p. 36), although it should be noted that details of the subjective system of evaluation, and what components of teaching were being looked at were not stated, and likely vary from district to district. The suggestion that Sass, Semykina, and Harris put forward is that although value-added models are flawed, they do provide a sense of direction and should not be ignored in terms of feedback, but likely should not be used as a method of evaluation (p. 36). Hill, Kapitula, and Umland (2011) agree with this assessment when the model omits prior years of education, and only looks at the current year, but note that many school districts may not have a thorough model, opting for simplicity as opposed to accuracy in the assessment model (p. 797).

Incentive-Based Evaluation

On its own, teacher evaluation simply remains a method to provide feedback to teachers, which is not the intent of teacher evaluation in many cases. Rather, a growing trend in education has been to include some sort of incentive for educators to perform well, according to whichever standards that the school district is using. These can contain flaws and be poor indicators of the performance of the individual teachers. One method commonly used to provide an incentive to a teacher to perform better is a pay-for-performance system, where the teachers that the system rates as more effective receive a higher pay than lower performing teachers.

According to Holt (2001), this method of an incentive can have significant negative of effects on the education system, both for the teachers and the students. In his view, teachers will not strive to be as effective as they can, because doing so will not necessarily guarantee them the appropriate reward. He claims that this is a misconception; however, he also asserts that there are many more factors besides the teacher that have an impact on the performance of the student (p. 312). Caillier (2010), agrees with Holt's conclusion that a pay-for-performance model for teachers is misguided, and agrees with Holt's reasoning, but provides several more. He argues that a pay-for-performance system is most effective when the task being judged is easily measurable for an individual, the organization provides the employee with a clear goal, and that the employee is highly motivated by monetary rewards. Caillier asserts that this is not the case with teachers, since multiple factors affect the student performance. Additionally, there is often not one clear goal, and teachers are public employees, who are less motivated by monetary rewards than private sector employees (p. 59-60). Liang and Akiba (2015) disagree with Callier and Holt, instead suggesting that if the pay can vary by as much as ten percent, it will motivate teachers to be more effective, but most focus on how a teacher teaches rather than how the students perform on standardized testing (p. 395). The impact that instituting the merit pay would have on students remains unclear, but it is assumed that it would positively impact students if implemented in a way that could accurately provide incentives to each teacher.

Results/Methods

Most of the literature on the subject of evaluating teachers on some method are qualitative, and not quantitative, and taken together, they highlight several major issues with the current methods used to evaluate teachers. This includes some methods that evaluate teachers based on the performance of their students, and the literature indicates that applying each policy in the correct scenario would make sense, but applying them in an incorrect scenario could lead to incorrect scoring of public school teachers and inaccurately assess student learning through a biased measure. The qualitative method has the advantage that numerical indicators of student success are not necessarily needed to be chosen, which can be difficult to do. However, the numerical methods tend to provide more insight into the current state of affairs but have to choose imperfect methods of measuring the performance of teachers. The literature indicates that the concept of having accountability in education is beneficial, at least as far as measures of teacher performance go, but the system of measuring this is flawed. The literature provides guides on how a teacher evaluation system might be structured to find one that aligns the incentives properly to produce beneficial results.

However, before placing too much emphasis on those results, a link needs to be drawn between student success and having the teachers evaluated, not simply between imperfect evaluation of teachers and the benefit for the school. The fact that the literature has not addressed this issue all that much is indicative of the fact that student success can be difficult to measure and will often need to be measured in a variety of ways. Furthermore, the methods used to evaluate students and teachers vary greatly from district to district, so that it can be difficult to collect enough data to do a true study on one type of evaluation system. Since the Rhode Island Department of Education standardized the teacher evaluation process across the state of Rhode Island, the evaluations are done in the same way across the state, with only subjective differences between schools. Therefore, the study can overcome some of the weaknesses seen in the literature and address whether or not evaluating teachers actually produces a benefit for the students, or if it simply is a program that appears beneficial to students when not examined in detail.

Conclusion

The literature discussed provides a backdrop that guides my study by illustrating a qualitative understanding of the issues and benefits of a variety of types of teacher evaluation, including value-added methods, high stakes testing, and subjective evaluation by a third party. The Rhode Island system of evaluation that I am considering includes a combination of all three, measuring the value added by a teacher by their ability to complete an SLO, or student learning objective, using student performance as an indication of teacher quality, and having an administrator or department chair do multiple observations of a classroom during the year. All of these have drawbacks if they are working separately, so it is possible that including all of these will reduce the overall inaccuracy, though the results of the study will determine if this is in fact the case. Furthermore, the limitation on data done in a few studies that have happened before have greatly reduced their reliability as evidence, so this study using data from a uniform, statewide evaluation system will be able to provide stronger evidence of its conclusions than most prior quantitative studies of the teacher evaluation system.

I am investigating multiple measures of student success, especially because the literature indicates that taking a narrow view and focusing in on any one measure of success eliminates the ability to measure student success in different ways and could be underrating teachers. Most significantly, the fact that some of the school districts are adopting evaluation processes that do not take into account concerns with the effectiveness of those systems raises concerns about the fairness of evaluation systems. The fact that teachers are not as well incented by money as private sector employees will not necessarily be observable in the study, but in the Rhode Island model, failure to perform well could lead to loss of job or loss of teaching license over time, thereby putting a different, negative incentive in place for Rhode Island teachers.

Rather than investigating the portions of the evaluation system separately, I investigated the Rhode Island evaluation system in general, which is a synthesis of most major evaluation methods. I will use statistical methods, which have only been able to be used when a whole state had similar evaluation methods in the past, such as Florida, but the evaluation methods were not necessarily consistent from district to district in that study. In Rhode Island, the evaluation system is mostly uniform across the state, allowing the data to much more comparable in my study than they were in the Florida study. The results of the study will indicate whether or not the Rhode Island system is effective in boosting student success, or if like the other evaluation systems that have been developed, it has its flaws as well. If that is the case, the theoretical backing of the literature and statistical results would lead me to make recommendations about how to strengthen the Rhode Island teacher evaluation system at the conclusion of the paper.

RHODE ISLAND TEACHER EVALUATION OVERVIEW

There are several components to the teacher evaluation system in Rhode Island, both objective and subjective. Teachers are evaluated based on professional practice and responsibilites in a subjective method by an administrator or other evaluator, based on rubrics, which can be found on the Rhode Island Department of Education website. The evaluation of professional practice includes both announced and unannounced classroom visits. Furthermore, teachers are evaluated by an objective method on professional growth and student learning objectives set at the beginning of the year. The evaluation results are communicated to a teacher who is getting evaluated through three conferences during the year with the evaluator (Rhode Island). These scores are combined into a final score, used to determine if a teacher is highly effective, effective, developing, or ineffective. Teachers scoring below effective have an improvement plan, while consistent underperformance could cause a teacher to lose his or her teaching certification. The requirement can be seen in Appendix L.

Teachers who were rated effective are evaluated every other year, while teachers who were rated highly effective are evaluated every three years, with the exception of untenured teachers or those with emergency certification, who are evaluated every year (Rhode Island). While the details can be very complex, a high level knowledge of the system is all that is necessary in order to carry out this statistical study, since it looks at students results in the state of Rhode Island and evaluates the entire evaluation system, rather than any one component of it. An example of one rubric used to evaluate teachers in the classroom is included in Appendix J.

METHODS

Rhode Island Study Overview

The purpose of the study was to compare student performance before and after the implementation of Rhode Island's teacher evaluation system in the 2012-2013 academic year. Since the intent was to evaluate the impact that the presence of the evaluation model had on student performance, and not to consider whether or not teachers rated effective or highly effective under the evaluation model had students who performed better than other teachers, I was not interested in looking at individual teacher ratings. Rather, I was interested in comparing data from across the state and determining what the impact of teacher on average student performance across the state was. To do this, I used data from prior to the implementation of the teacher evaluation model and data from after the implementation and determined whether or not there was a difference between student performance before and after teacher evaluation was used that was not able to be attributed to anything else. As the evaluation does not apply to private school teachers, only public high schools were considered. Futhermore, the study was limited to traditional high schools, excluding vocational schools and charter schools due to inherent differences between these types of schools and traditional public high schools.

Data Collection

Ideally, the data used for the study would have been information about every single public high school student in Rhode Island, as well as information about the schools that they attended, including if they switched schools during high school. However, due to privacy concerns these data were not available. Instead, the data used for the study came from Rhode Island Infoworks, a database maintained by the Rhode Island Department of Education, or RIDE. The data used for the study were averages for each high school in each academic year beginning with 2008-2009 and ending with 2014-2015, the most recent complete academic year. It is important to note that this method is not ideal since this weights all schools equally, which means that a student attending a smaller school would have a larger impact on the final results than a student attending a larger school. However, since the number of students attending each school in each year was not available on Infoworks, each individual school was used as a single source of data. It is worth noticing that all school districts report data separately to RIDE, and that this could possibly lead to a lack of uniformity in how data were reported. Since the data considered are very objective, this was not seen as a big concern in the study.

The types of information considered in the study included information about standardized tests, including both the Scholastic Aptitude Test, or SAT, commonly taken by students prior to attending college, and the New England Common Assessment Program, or NECAP test, which was mandatory for students to take during the academic years included in the study. The Advanced Placement, or AP Exams, which can give students college credit if passed, were also considered, although they only indicate performance of the top students in a school. In addition, characteristics of the students at the schools, including such things as eligibility for subsidized lunch, and participation rates in Special Education and English Language Learner, or ELL Programs were also included. Furthermore, teacher certifications, and student attendance and graduation information were considered, as were annual suspensions at each school.

One exception to this was student characteristics, which were not available in 2008-2009. Additionally, suspension, SAT, and NECAP math, reading, and writing data, were not available in 2014-2015 at the time of data collection. At the time of writing, SAT data has recently been provided for 2014-2015, but was not added back into the study. Since these data were not available, it was considered missing data, which will be further discussed. Additionally, Infoworks had information on school accountability, adequate yearly progress, and the Partnership for Assessment of Readiness for College and Careers, or PARCC test, but school accountability included three years of data, while the others included only a single year of data, so there simply were not enough data to use these in the study. Thus, my sample size was 353 different points, each one representing one Rhode Island public high school in one particular academic year. Details about the data may be seen in Appendix K.

Data Modification

After collecting the data and joining it all into one large data table, listing the data by academic year and school, several changes needed to be made before analyzing the data. First of all, two public high schools in Rhode Island had to be eliminated from the data set due to issues with data. The first school to be eliminated from the study, Hope High School, was removed simply because it had insufficient data to be included in the study. The school had been closed and reopened in the 2011-2012 academic year, which only allowed for a single of year of data from the school prior to the implementation of the teacher evaluation. Therefore, Hope High School was not included in the study. Block Island School was also not included in the study, since it is a K-12 school, and not just a high school, and there was no method available to split the data between the high school students and the elementary and middle school students attending the school. Additionally, some method of indicating the presence of the teacher evaluation model was needed. I therefore defined a new variable in the dataset, TeacherEval, set equal to 1 in academic years when teacher evaluation was present, which was all years from 2012-2013 until the present. All prior years were assigned a 0, since the teacher evaluation system was not in place at the time.

Statistical Analaysis

The initial statistical analysis that occurred used the TeacherEval variable as the dependent variable and used the other variables as predictors. I built a decision tree from these inputs in SAS Enterprise Miner using default settings, using all other variables as predictor, except for the school and the year. I was interested in overall averages, not school by school differences, and since TeacherEval was defined based on the year, the year perfectly predicted the variable. Thus, I also wanted to run logistic regressions models and compare the results in order to find the best model. However, with the presence of missing data, this was initially impossible. One solution would have been simply to eliminate the variables that had missing values as predictors, but to do so I would need to eliminate a significant amount of the data. Instead, I chose to impute data with a decision tree to fit the missing values, using all variables, including the school but excluding the year, as a predictor. While not as accurate as having actual data in all cases, this was a reasonably accurate method to fill the missing data and to be able to use all of the predictors without dramatically shrinking the sample size.

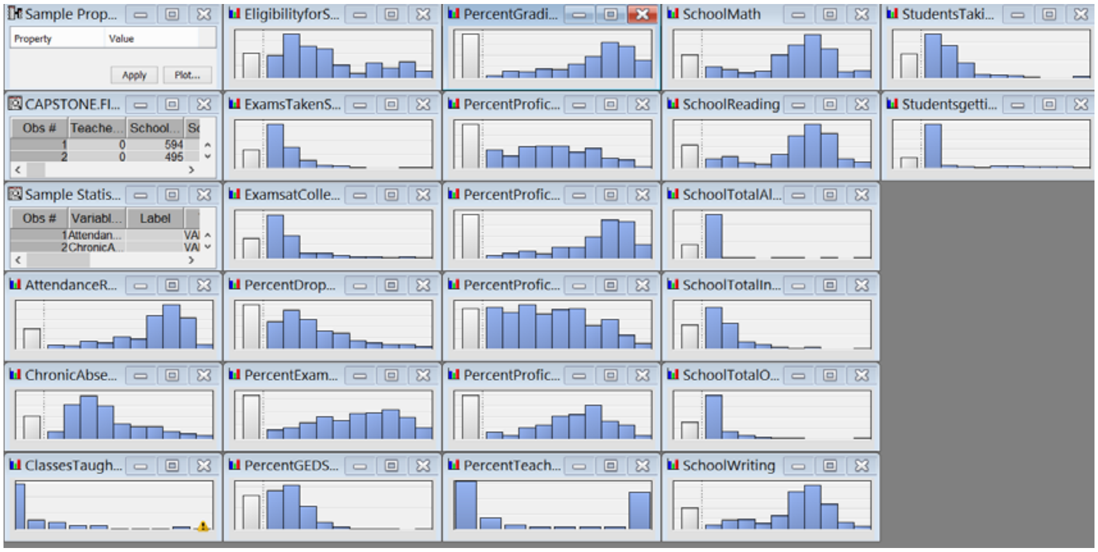
I then ran a stepwise logistic regression model to determine what other factors TeacherEval was correlated with, and then out of those, I tested to see if quadratic terms or interaction terms between the factors were correlated with TeacherEval. I used a significance level of 0.05 to determine whether or include a given variable or remove it from the model at each step. Finally, I compared all of the models using the misclassification rate to determine which model was better. The misclassification rate was the number of cases in which the model inaccurately determined whether or not the data included teacher evaluation or not; I chose the model where it was lowest. If the difference between the misclassification rate for the best model and the next best model was small, I picked the next best model if it was less complex than the one with a slightly lower average squared error. I also used SAS 9.4 to verify the assumptions for regression.

I then repeated this process for each subject area of the SAT: math, reading, and writing, as well the four year graduation rate, and each subject area of the NECAP: math, reading, writing, and science. The purpose of this was to see if the presence of teacher evaluation led to any meaningful differences in any individual metrics for student success. However, there were several differences from the previous process. When I ran all of the models, I chose to not use some variables that would be very correlated with each other. I did not use the dropout rate or the GED rate as predictors for the four year graduation rate since there is a strong relationship; all students must graduate, get a GED, or drop out of school. Additionally, the NECAP math test was not used as a predictor for the SAT math test and vice-versa, and likewise for each of reading and writing, since correlation between these tests was expected; students with an aptitude for a subject were likely to consistently be stronger in that area.

Additionally, while TeacherEval is a binary variable, and only takes on the values 0 and 1, the other variables that I looked as a dependent variable were not and therefore, I ran linear regressions rather than logistic regressions. Furthermore, misclassification rate is not applicable when the dependent variable is not binary, so the model with the minimal average squared error was chosen as the best model in each case, or one with an average squared error only slightly higher than the best model that was also simpler than the other model. If a regression had many terms, most of which were significantly insignificant, the model was rejected due to most of it reflecting random effects. The best model in each variable was used in order to determine the results.

Verification of Assumptions for Regression

Looking at the histograms of all the variables below, it can be seen that not all variables are normally distributed. However, based on a visual examination, the measure of student success, such as the four year graduation rate, SAT scores, and percent proficient on NECAP tests can be seen as roughly normal since they are somewhat symmetric and unimodal.



Obs #	Teache...	School...	Sc
1	0	594	
2	0	495	

Obs #	Variabl...	Label	
1	Attendan...		VA
2	ChronicA...		VA

Figure 1 - Distribution of Variables

Furthermore, since averages are being predicted, and the number of data points is 353, the central limit theorem would imply that the average values would be much closer to being normally distributed. The data points are definitely not completely independent, being based off of schools in consecutive academic years, so individual students can affect the data for a total of four years. However, the schools are somewhat independent of each other, and rather than sampling schools, I used the entire population of Rhode Island public high schools that were not vocational or charter schools that had high school data for the entire period under consideration.

The fit diagnostics for the regression predicting TeacherEval is seen in Appendix A. The top left graph is a residual plot, and shows that there are no real relationships between the residuals and that they are random. The graph in the second row and first column is a Q-Q plot and the graph following the straight line would indicate that residuals are normally distributed. This is true in the middle, but schools on either extreme do not quite fit the model. Additionally, the histogram on the bottom left indicates that the data is somewhat normally distributed, but not perfectly. This indicates that the regression somewhat fits the assumptions in this case.

The fit diagnostics for the regression predicting the SAT Mathematics Score is seen in Appendix B. The residual plot indicates that there is no relationship between residuals and that they are random. The Q-Q plot indicates normality of residuals since they mostly fall in a straight line, and the histogram indicates that the data are roughly normal. The fit diagnostics for the regression predicting the SAT Reading Score is seen in Appendix C. The residual plot indicates that there is no relationship between residuals and that they are random. The Q-Q plot indicates normality of residuals since they mostly fall in a straight line, and the histogram indicates that the data are roughly normal. The fit diagnostics for the regression predicting the SAT Writing Score is seen in Appendix D. The residual plot indicates that there is no relationship between residuals and that they are random. The Q-Q plot indicates normality of residuals since they mostly fall in a straight line, and the histogram indicates that the data are roughly normal.

The fit diagnostics for the regression predicting the percent of students proficient on the NECAP Mathematics test is seen in Appendix E. The residual plot indicates that there is no relationship between residuals and that they are random. The Q-Q plot indicates normality of residuals since they mostly fall in a straight line, and the histogram indicates that the data are roughly normal. However, the assumption of normal residuals predicts worse than actual performance by the top performers. The fit diagnostics for the regression predicting the percent of students proficient on the NECAP Reading test is seen in Appendix F. The residual plot indicates that there is no relationship between residuals and that they are random. The Q-Q plot indicates normality of residuals since they mostly fall in a straight line, and the histogram indicates that the data are roughly normal. The fit diagnostics for the regression predicting the percent of students proficient on the NECAP Writing test is seen in Appendix G. The residual plot indicates that there is no relationship between residuals and that they are random. The Q-Q plot indicates normality of residuals since they mostly fall in a straight line, and the histogram indicates that the data are roughly normal. However, the assumption of normal residuals predicts better than actual performance by the top performers The fit diagnostics for the regression predicting the percent of students proficient on the NECAP Science test is seen in Appendix H. The residual plot indicates that there is no relationship between residuals and that they are random. The Q-Q plot indicates normality of residuals since they mostly fall in a straight line, and the histogram indicates that the data are roughly normal.

The fit diagnostics for the regression predicting the four year graduation rate is seen in Appendix I. The residual plot indicates that there is no relationship between residuals and that they are random. The Q-Q plot indicates that residuals are close to being normal, they are not quite normal since they mostly fall in a straight line, but show significant fluctuation. The histogram indicates that the data are not roughly normal, indicating that regression may not be as accurate for predicting the graduation rate.

<u>INITIAL OBSERVATIONS</u>

Before looking at the statistical results, it is important to understand how students having been performing in Rhode Island since the fall of 2008, where the earliest data comes from. Since the fall of 2008, the average four year graduation rate has increased from 78% to 84%, as can be seen in Figure 2.

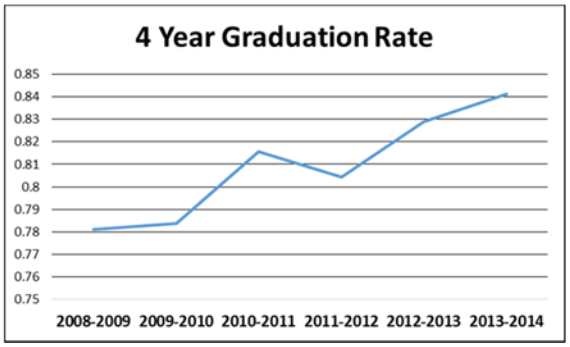


Figure 2 - Graduation Rate

This is an upward trend that began prior to the implementation of teacher evaluation, although there was a particularly large spike from the 2011-2012 to 2012-2013 school years. However, the graduation rate in 2011-2012 was abnormally low, so this could be due to the implementation of teacher evaluation, a recovery from the dip in the previous year, or even just a continuation of a previous trend. This paper will, among other things, look at whether or not teacher evaluation is responsible for these changes.

Looking at the SAT Scores over the same time period Figures 3 and 4 below, a downward trend becomes visible.

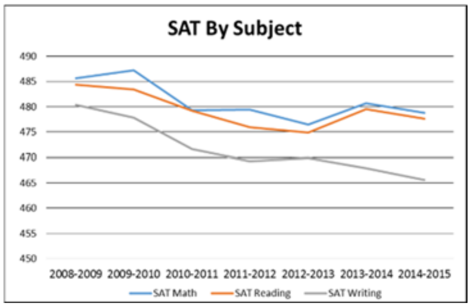


Figure 3 - SAT Subject Scores

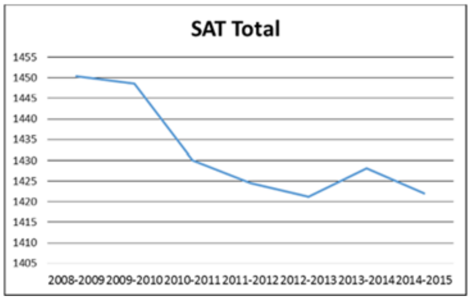


Figure 4 - SAT Total Score

Looking at the percent of students proficient on the NECAP Exam from 2008-2015, a positive trend is visible in every subject area, although this may be attributable to the NECAP test becoming a graduation requirement in the 2011-2012 school year. This is visible in Figure 5.

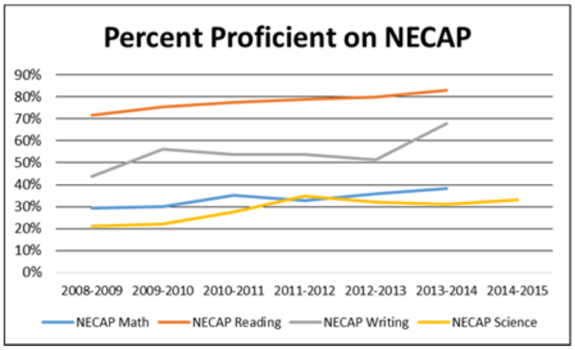


Figure 5 - Percent of Students Proficient on NECAP by Subject

Finally, changing characteristics of Rhode Island students could have impacted student performance between 2008 and 2015. Over that time period, the percent of students eligible for subsidized lunch increased from 34 percent to 38 percent, as can be seen in Figure 6.

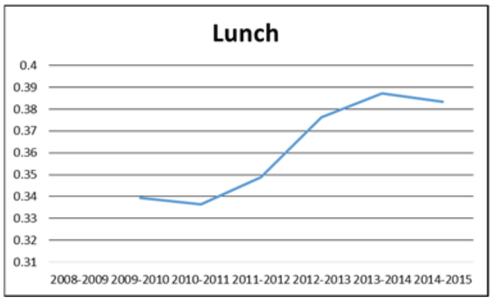
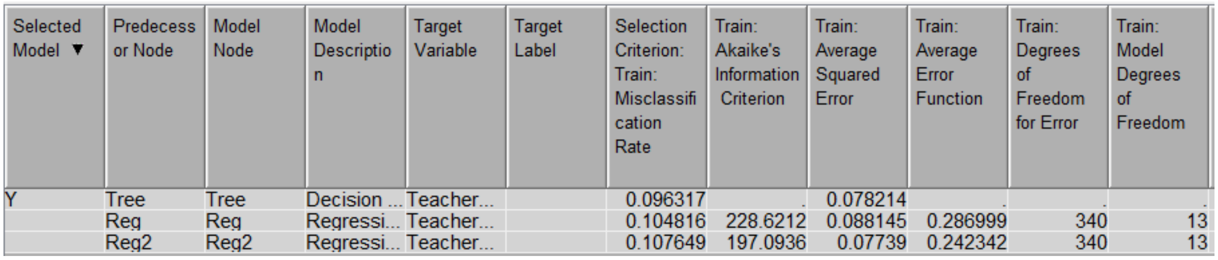


Figure 6 - Percent of Students Eligible for Subsidized Lunch

More students being eligible for subsidized lunch indicates students of lower socioeconomic status, which has been reliably shown to have an impact on student performance. Whether these trends are related to or separate from teacher evaluation will be established by the statistical results.

RESULTS

Investigating the different models that predicted whether teacher evaluation was present or not, I selected the one with the lowest average squared error. As can be seen below in Figure 7, the decision tree has the lowest misclassification rate, 9.63%, while the regression without interaction had one of 10.48%, and the regression with interaction terms had a misclassification rate 10.76%.



Selected Model ▼	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Train: Misclassification Rate	Train: Akaike's Information Criterion	Train: Average Squared Error	Train: Average Error Function	Train: Degrees of Freedom for Error	Train: Model Degrees of Freedom
Y	Tree	Tree	Decision ...	Teacher...		0.096317	.	0.078214	.	.	.
	Reg	Reg	Regressi...	Teacher...		0.104816	228.6212	0.088145	0.286999	340	13
	Reg2	Reg2	Regressi...	Teacher...		0.107649	197.0936	0.07739	0.242342	340	13

Figure 7 - Selection of Best Model for TeacherEval

Therefore, I concluded that the decision tree best indicated which variables were most related to the presence of teacher evaluation, since it only incorrectly classified 9.63% of schools and academic years as having teacher evaluation when they did not, or the other way around.

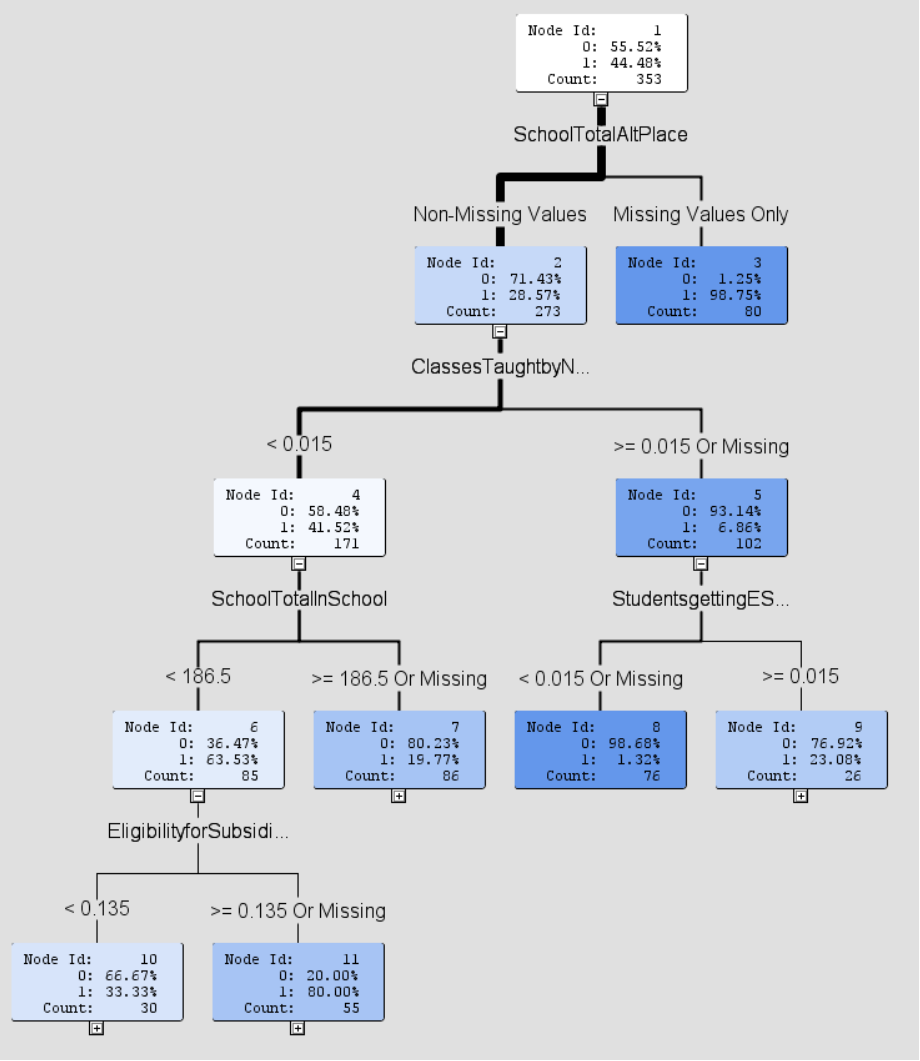


Figure 8 - Decision Tree Model For TeacherEval

As can be seen in Figure 8, the variables that are overall mostly closely related to the presence of teacher evaluation are alternate placements for suspensions, classes taught by not highly qualified teachers, in school suspensions, students getting ESL, and eligibility for subsidized lunch. Although classes taught by not highly qualified teachers appears to indicate the number of courses taught by teachers who had not performed well on the evaluations, it is completely unrelated to that issue. Rather, this simply indicates the number of teachers on emergency certification, which is given to teachers when schools need a teacher quickly, so that the teacher can work towards their full certification while teaching. Therefore, the best predictors of teacher evaluation are related to student discipline, English as a Second Language, teacher certification, and socioeconomic status of students. The interesting conclusion that this reveals is that none of the measures of student success seem to show any relationship to the teacher evaluation model when considered overall. Therefore, I went and looked at each individual measure of student success to see if the presence of teacher evaluation had caused changes in any individual measure.

I began this process by reviewing the SAT scores by subject area. For the model with the lowest average square error, the regression with interaction terms had the lowest average squared error, 79.56, while the regression without the interaction terms had an average squared error of 84.06. However, including the terms that were interacting that had been removed by the stepwise regression process decreased the average squared error of the model with interaction to 74.73, while most variables, including interaction terms that were significant in the stepwise model, became statistically insignificant. Therefore, the regression without interaction was chosen since the model with interaction became a poor model when the original variables were considered as well. The results of the regression are in Figure 9.

Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-17.8408	9.5807	-1.86	0.0635
IMP_PercentProficientScienceScho	1	40.448	5.3695	7.53	<.0001
IMP_PercentTeacherswEmergencyCer	1	119.7	42.8657	2.79	0.0056
IMP_SchoolReading	1	0.584	0.0709	8.24	<.0001
IMP_SchoolWriting	1	0.4189	0.0714	5.87	<.0001
IMP_StudentsgettingESLSchool	1	88.4764	13.4581	6.57	<.0001
IMP_StudentsgettingSpecialEdScho	1	39.7502	12.8438	3.09	0.0022
TeacherEval 0	1	1.4742	0.6107	2.41	0.0164

Figure 9 - Regression Model for SAT Mathematics Score

While the SAT Mathematics score had a significant relationship to the percent of students in the school proficient on the science NECAP test, and the SAT Reading and Writing test, and the percent of teachers with emergency certification, and the percent of students getting ESL and Special education, in addition to these, there is a strong relationship between the presence of teacher evaluation and school average SAT Mathematics scores. While this relationship is statistically significant, however, it is only a difference of 1.47 points on the SAT scale of 200-800 for each subject.

For the SAT Reading test, the results are not as clear as with the SAT Mathematics test. The decision tree has the highest average squared error of all the models, at 123.87, and was therefore not even considered to possibly be the best model. The regression without any interactions had an average squared error of 43.6, while the regression with interaction terms had an average squared error of 41.6. Including the model with interaction terms, as well as the terms that were interacting, the average squared error dropped to 39.1. However, this model also had many insignificant terms, including interaction terms that had been significant in the stepwise regression with interactions. Because of this, the interaction model is not ideal, so I used the simplest model, the regression without any interactions, since its average squared error was only 4.5 SAT points higher than the best model out of a scale of 200 points to 800 points. The simplest regression can be seen in Figure 10.

Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	23.3094	6.1387	3.8	0.0002
IMP_PercentProficientScienceScho	1	-9.3604	4.0164	-2.33	0.0204
IMP_SchoolMath	1	0.293	0.0366	8	<.0001
IMP_SchoolWriting	1	0.6735	0.0379	17.77	<.0001
IMP_StudentsTakingExamSchool	1	0.0213	0.006	3.54	0.0005
IMP_StudentsgettingESLSchool	1	-53.4721	9.9724	-5.36	<.0001
TeacherEval 0	1	-1.8134	0.4291	-4.23	<.0001

Figure 10 - Regression Model for SAT Reading Score

It can be seen that the percent of students proficient on the Science NECAP Exam, the SAT Math, SAT Writing, Number of Students Taking AP Exams, and percent of students getting ESL have a statistically significant effect on the SAT Reading test, and so does the teacher evaluation model. The statistically significant effect from the teacher evaluation model is a decrease of 1.81 points on the SAT Reading test, which is a very counterintuitive result. However, this is on a scale of 200-800 points, so it is in practice very small.

For the SAT Writing test, the regression without interaction had the lowest average square error of all models considered, with 47.50. Although the regression with interaction terms had an average square error only slightly higher, at 47.67, once the terms that are interacting are added back into the model, the average square error decreases to 45.39. However, in addition, many variables become insignificant, including interaction terms that were significant in the stepwise regression model with interaction, so this model was not considered a good model. The decision tree, with an average square error of 188.74, was eliminated from consideration due to its extremely high average square error. The regression without interaction terms was thus chosen as the best model and can be seen in Figure 11.

Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	7.3622	5.1326	1.43	0.1525
IMP_PercentTeacherswEmergencyCer	1	-92.9352	31.5571	-2.94	0.0035
IMP_SchoolMath	1	0.2206	0.0352	6.28	<.0001
IMP_SchoolReading	1	0.7598	0.0365	20.8	<.0001
IMP_StudentsTakingExamSchool	1	-0.0232	0.00647	-3.59	0.0004
IMP_StudentsgettingSpecialEdScho	1	-21.8156	9.7401	-2.24	0.0258
TeacherEval 0	1	1.6722	0.4261	3.92	0.0001

Figure 11 - Regression Model for SAT Writing Score

The percent of teachers with emergency certifications, SAT Math score, SAT Reading score, the number of students taking AP Exams, and the percent of students getting special education have a statistically significant effect. However, it is important to note that the teacher evaluation model does as well, and that once again, the statistically significant difference is small; here it is 1.67 points.

Looking at the percent of students proficient in the Math NECAP test, the decision tree model has an average square error of 0.003916, while the regression without any interaction terms has an average square error of 0.002759. The regression with interaction terms has average square error of 0.002288, but when adding the terms that are interacting back into the model, this decreases to 0.002159, while many terms, including interaction terms that were significant in the previous stepwise regression model, become insignificant. Thus, I picked the regression with no interaction as the best model, which can be seen below in Figure 12.

Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-0.3868	0.0401	-9.66	<.0001
IMP_ExamsatCollegeMasterySchool	1	0.000267	0.000043	6.17	<.0001
IMP_PercentProficientScienceScho	1	0.5004	0.0329	15.22	<.0001
IMP_PercentProficientWritingScho	1	0.1634	0.0274	5.97	<.0001
IMP_SchoolReading	1	0.000998	0.0001	9.95	<.0001
IMP_SchoolTotalInSchool	1	-0.00004	0.000014	-2.72	0.007

Figure 12 - Regression Model for Percent of Students Proficient on NECAP Mathematics

The variables that are related to the percent of students proficient on the Math NECAP are the number of AP exams passed, the percent of students proficient on the science and writing NECAP, the SAT Reading test score, and the number of in-school suspensions. None of these are related at all to teacher evaluation, so this demonstrates that teacher evaluation did not significantly change the percentage of students proficient in the Math NECAP test.

Looking at the percent of students proficient on the NECAP reading test, the decision tree had the highest average square error of 0.00299, and is therefore definitely not the best model. The regression without any interaction had an average square error of 0.002319, and the stepwise regression with interaction terms had an average square error of 0.001701. When the main effects whose interactions were significant were added back into the model, the average square error fell to 0.001402, but in the process many previously significant terms became statistically insignificant. Therefore, the regression without interaction, which still had a very small average square error, was chosen as the best model, and can be seen below in Figure 13.

Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	0.0735	0.1038	0.71	0.4799
IMP_AttendanceRateSchool	1	0.3826	0.1216	3.15	0.0018
IMP_PercentGradin4YearsSchool	1	0.2487	0.0498	4.99	<.0001
IMP_PercentProficientScienceScho	1	0.1631	0.0292	5.59	<.0001
IMP_PercentProficientWritingScho	1	0.3136	0.0277	11.32	<.0001
IMP_PercentTeacherswEmergencyCer	1	-0.5192	0.238	-2.18	0.0301
IMP_SchoolTotalOutofSchool	1	0.000026	7.66E-06	3.39	0.0008
IMP_StudentsgettingESLSchool	1	-0.5189	0.0723	-7.17	<.0001
IMP_StudentsgettingSpecialEdScho	1	-0.2984	0.0797	-3.75	0.0002
TeacherEval 0	1	-0.00848	0.00353	-2.4	0.017

Figure 13 - Regression Model for Percent of Students Proficient on NECAP Reading

While the attendance rate, four year graduation rate, percent of students proficient on Science and Writing NECAP Exams, Percent of teacher with emergency certification, the number of out of school suspensions, and the percent of students getting ESL and Special education all had significant results, it is important to note that the teacher evaluation system did as well. Due to the presence of teacher evaluation, on average, 0.848% fewer students were proficient on the Reading NECAP, a number that, although statistically significant, is very small in practice.

When considering the percent of students proficient on the Writing NECAP, the decision tree had the largest average square error of 0.007684, while the regression without any interaction had a lower average square error of 0.007049. The regression with interaction terms had an even lower average square error of 0.006850, which decreased further to 0.006302 when the main effects that had significant interactions were added back into the model. However, this happened while most terms in the model became statistically insignificant. Therefore, I selected the regression model without any interactions as the best model, and it can be seen in Figure 14.

Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	0.042	0.0768	0.55	0.5848
IMP_PercentProficientMathSchool	1	0.4434	0.086	5.15	<.0001
IMP_PercentProficientReadingScho	1	0.9695	0.0694	13.97	<.0001
IMP_PercentProficientScienceScho	1	-0.2634	0.0727	-3.62	0.0003
IMP_PercentTeacherswEmergencyCer	1	0.8389	0.3935	2.13	0.0339
IMP_SchoolReading	1	-0.00068	0.000198	-3.43	0.0007

Figure 14 - Regression Model for Percent of Students Proficient on NECAP Writing

This model indicates that the percent of students proficient on the math, reading, and science portions of the NECAP, as well as the SAT Reading test score and the percent of teachers with emergency certification. However, there is no impact on the SAT writing score from the teacher evaluation model.

Considering the final subject area of the NECAP, science, the decision tree model had the highest average square error, 0.006041, it was not even considered as possibly being the best model. The regression with no interactions terms had an average square error of 0.004952. The regression with interaction terms had an average square error of 0.004272, which was lower, and when the terms whose interactions were significant in the model that included interaction terms are added back in, the average square error drops to 0.003464. However, when they are included, nearly every term in the regression model becomes insignificant, including interactions that the stepwise regression had concluded were significant. Therefore, I selected the regression without interaction since it did not result in the complications that arose with interaction, but the average square error was still very close. The model can be seen in Figure 15.

Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	1.0199	0.2173	4.69	<.0001
IMP_ClassesTaughtbyNotHighlyQual	1	-0.3787	0.1298	-2.92	0.0038
IMP_EligibilityforSubsidizedLunc	1	-0.134	0.037	-3.62	0.0003
IMP_ExamsatCollegeMasterySchool	1	0.00022	0.000061	3.58	0.0004
IMP_PercentGradin4YearsSchool	1	-0.1982	0.0693	-2.86	0.0045
IMP_PercentProficientMathSchool	1	0.645	0.0617	10.45	<.0001
IMP_PercentProficientReadingScho	1	0.3468	0.0573	6.05	<.0001
IMP_PercentProficientWritingScho	1	-0.1718	0.0475	-3.62	0.0003
IMP_SchoolMath	1	0.00121	0.000394	3.07	0.0024
IMP_SchoolReading	1	-0.00144	0.000379	-3.81	0.0002
IMP_TeacherStudentRatioAllTeache	1	-16.9295	3.7679	-4.49	<.0001

Figure 15 - Regression Model for Percent of Students Proficient on NECAP Science

The variables that have a significant relationship to the percent of students proficient in the NECAP Science test were the number of classes taught by not highly qualified teachers, the percent of students eligible for subsidized lunch, the number of AP exams passed by students at the school, the four year graduation rate, the percent of students proficient on each of the NECAP math, reading, and writing tests, the SAT Math and Reading test scores, and the student-teacher ratio. The teacher evaluation system did not have a significant effect on the percent of students proficient on the science NECAP test.

Finally, looking at the models predicting the four year graduation rate, the regression with no interactions has the highest average square error of 0.002893, while adding the interactions decreases the average square error to 0.002733. The decision tree has an average square error of 0.002633, while adding the main effects back into the regression with interactions decreases the average square error to 0.002522. However, that regression has most terms insignificant, including interactions there were significant prior to adding the main effects back into the model. Thus, I selected the decision tree, which had the second lowest average square error, by a little bit, but did not have the complications of the regression with the lowest error as the best model. It can be seen in Figure 16.

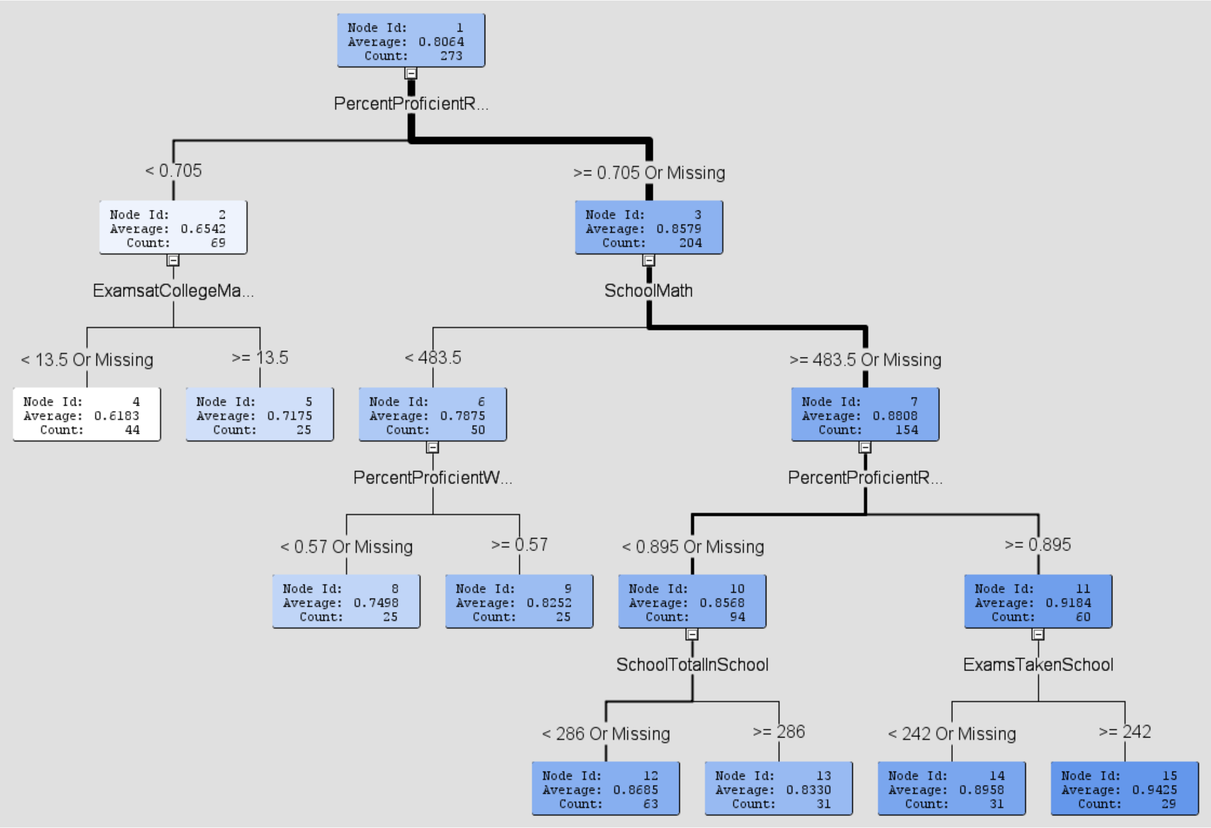


Figure 16 - Decision Tree Model for Four Year Graduation Rate

For those schools with fewer than 70.5 percent of students proficient on the reading NECAP in a given academic year, as well as fewer than 13.5 AP Exams passed or no information about the number of AP Exams passed, the model would predict that the graduation rate would be 61.83 percent. If the percent of students proficient on the Writing NECAP was still below 70.5 percent, but at least 13.5 AP Exams were passed by the students of the school, then the model predicts an expected graduation rate of 71.75%. If at least 70.5 percent of students were proficient on the Reading NECAP, and the average SAT Math score was less than 483.5, the model predicts an expected graduation rate of 78.75%. If instead, the score is not known or is at least 483.5, then the predicted graduation rate is 88.08%. The model continues to branch out most of these into further cases, but it is important to note that the only variables that it uses to predict the graduation rate are the percent of students proficient on the NECAP Reading and Writing tests, the average SAT Math score, the number of AP Exams taken and the number of AP Exams passed, and the number of in school suspensions. None of these predictors are the teacher evaluation system.

It is interesting to note that the cases where the regression did not quite meet the underlying assumptions do not affect the validity of the conclusions, as in both cases, TeacherEval and the graduation rate, the decision tree was the better model. Thus all regressions that were selected as the best model meet all assumptions for regression and are statistically sound.

CONCLUSIONS AND IMPLICATIONS

The statistical analysis reveals no overall relationship between the various measures of student success and the Rhode Island teacher evaluation system. When looking at these measures individually to determine if there is a significant effect on any student success measure by itself, no effect is found in most cases, while only a small effect is found in all of the other cases. No relationship is found between the teacher evaluation system and the four year graduation rate, nor is a relationship found between teacher evaluation and the percent of students proficient on the NECAP Mathematics, Writing, and Science tests. There is a statistically significant change in the percent of students proficient on the NECAP Reading test, a decrease of 0.848 percent of students, which is a small change in terms of magnitude. Likewise, teacher evaluation had a statistically significant effect on the average SAT score in each subject area, with an increase of 1.47 points on the mathematics test, a decrease of 1.81 points on the reading test, and an increase of 1.67 points on the writing test. These changes are very small in magnitude, especially given that SAT score on any given subject area ranges from 200-800.

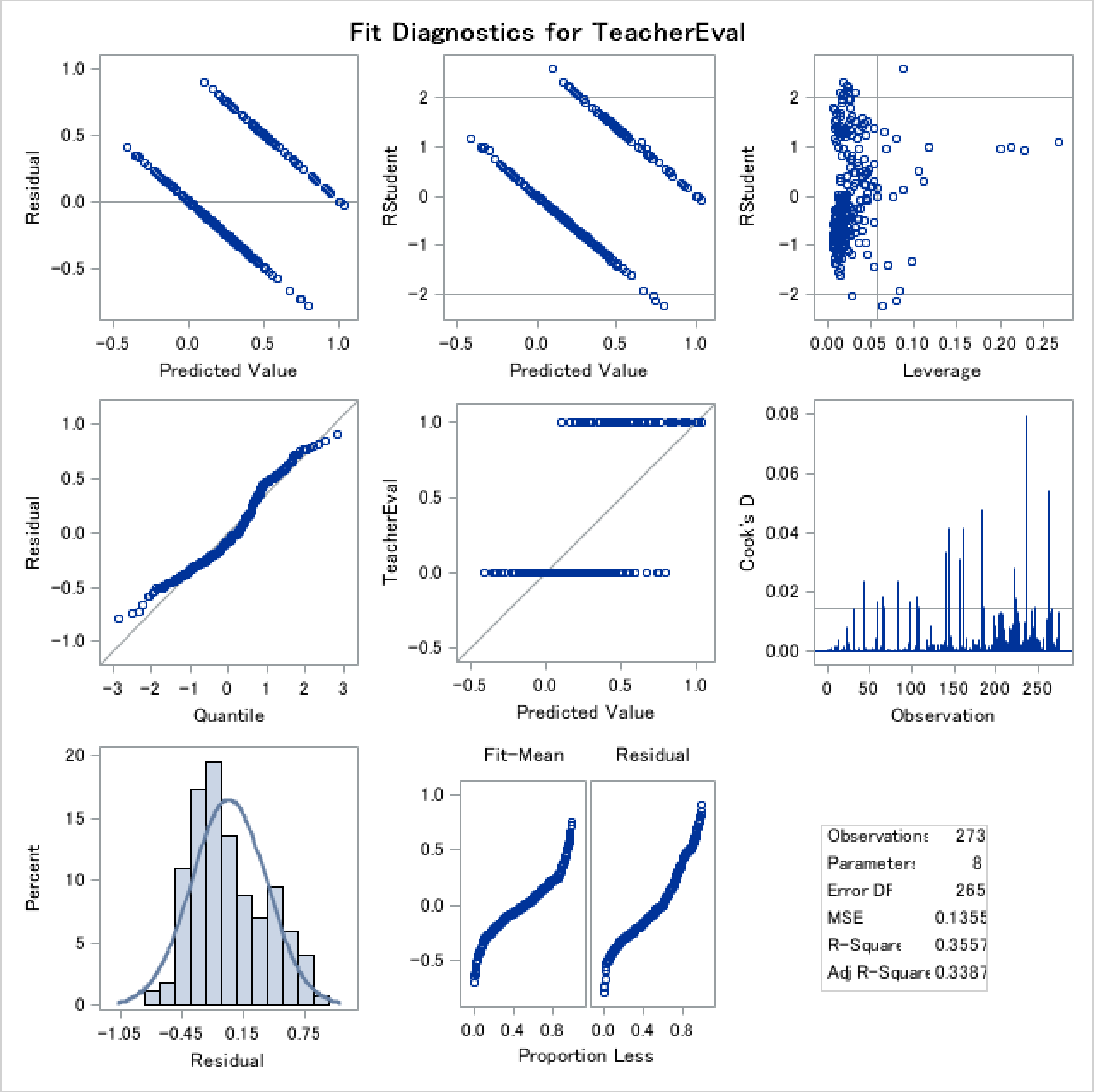
Since all changes due to the teacher evaluation model are either nonexistent or extremely small, this would imply that the teacher evaluation has not been successful in its purpose of improving student performance in Rhode Island. Furthermore, in the results, the percent of students in special education and the percent of students in English as a Second Language had significant negative impacts on students' success. Other issues that consistently were related was performance on different standardized tests, which partially predicted each other, and disciplinary issues which would decrease student performance.

The results indicate that in order to achieve the intended benefit for students at a school, average performance can be increased by increasing support for special education and ESL, rather than spending money and time on evaluating teachers, which does not produce the desired benefit. Not only would this support the students who use these programs, this would better enable classroom teachers to turn some of the extra attention they need to give students in special education and ESL to other students, if special education and ESL are themselves better supported than currently. Therefore, supporting those programs would benefit both students who participate in them and students who do not. Overall, the statistical analysis demonstrates than this would increase student performance by addressing factors that consistently decrease overall student performance, and have a much larger impact that evaluating teachers, which has no impact on student success in Rhode Island.

In the future, it would be useful to repeat the study with individual student data to increase accuracy, determine if the new PARCC test is useful as a predictor of student success in other areas, and to determine if the small changes made in the details but not the overall structure of the teacher evaluation system in the 2015-2016 academic year caused the teacher evaluation system to benefit students. This study only considered 2008-2015, and therefore did not incorporate that change.

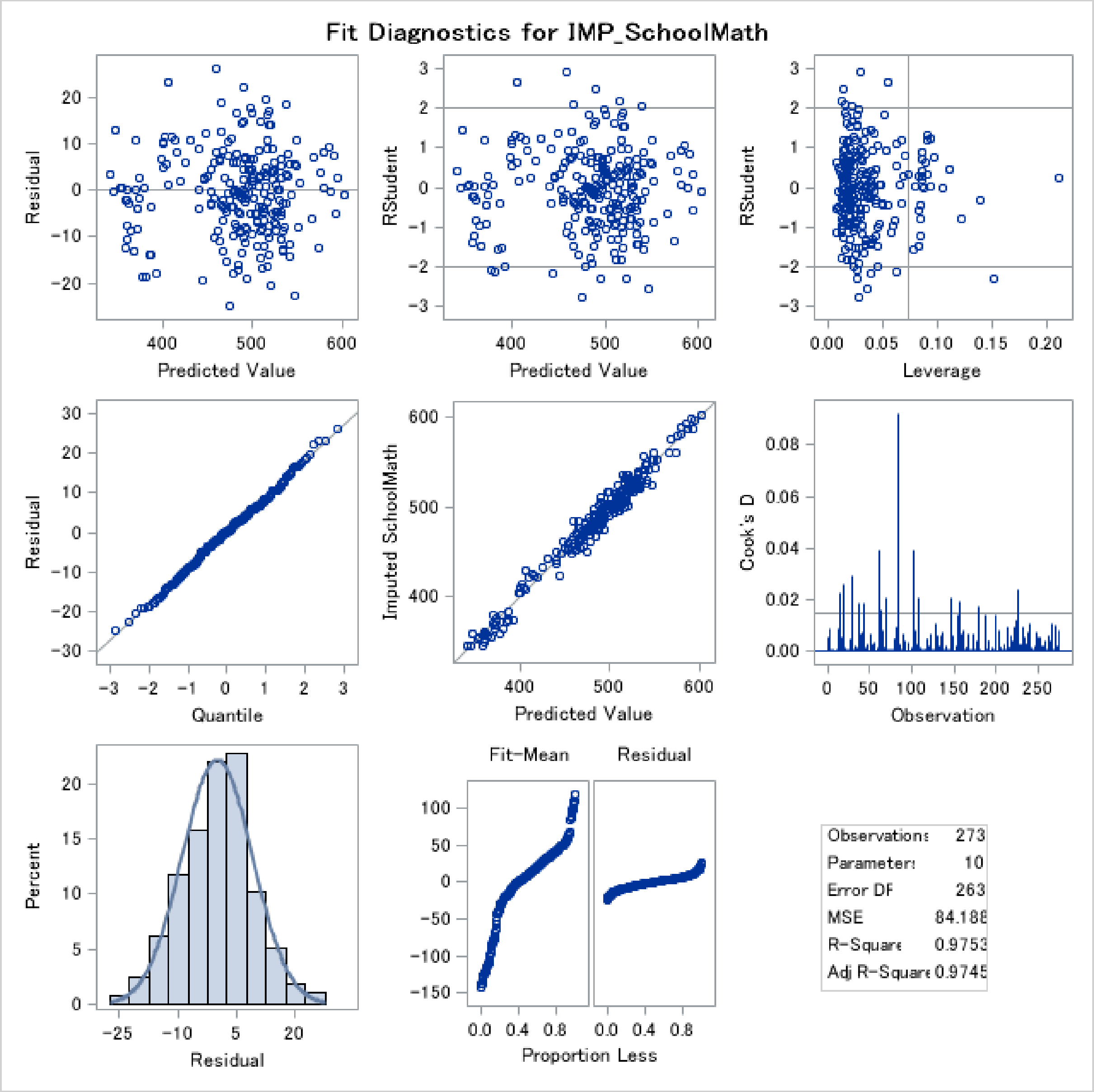
APPENDICES

Appendix A: Regression Assumptions TeacherEval



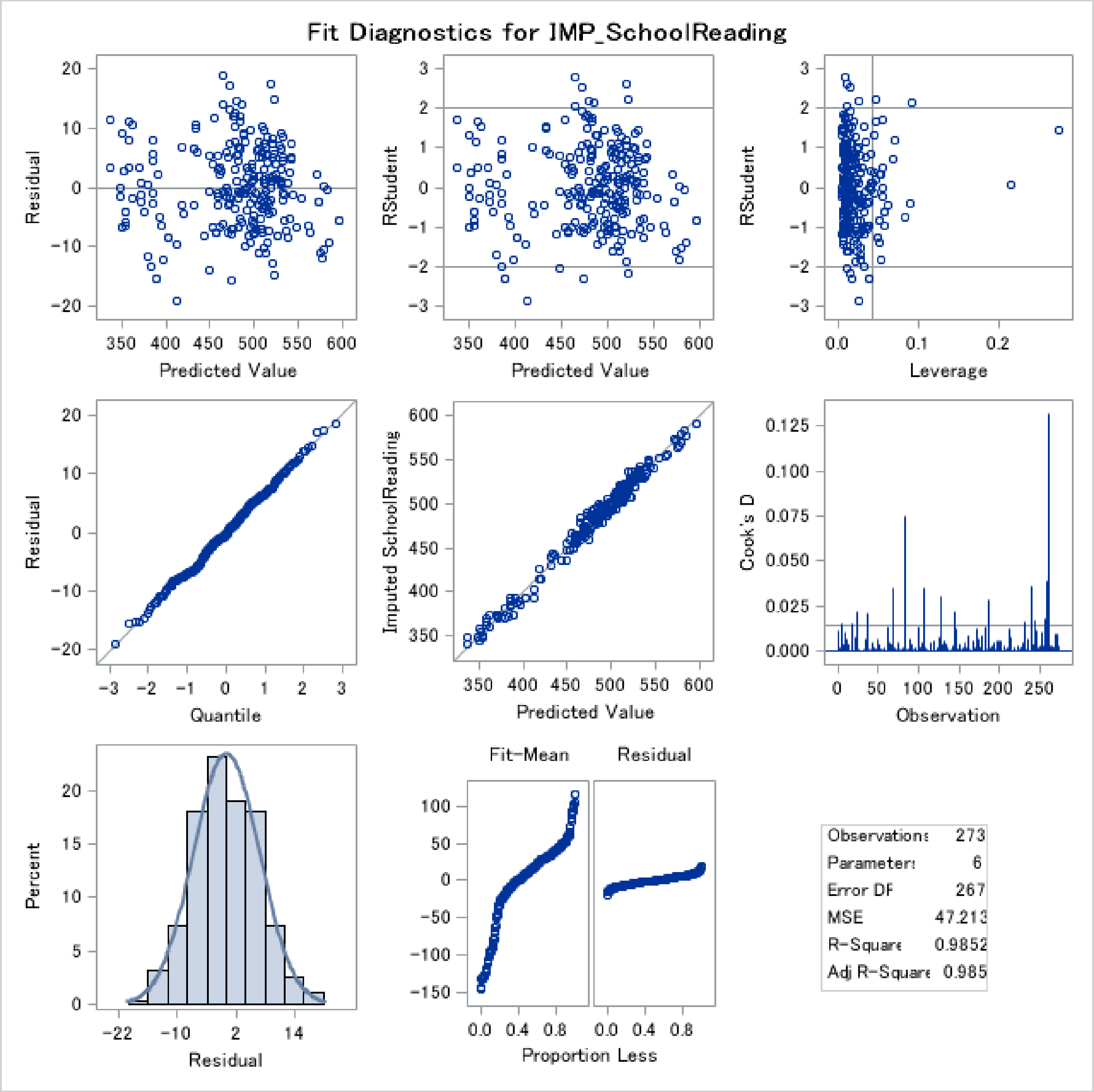
Observations	273
Parameters	8
Error DF	265
MSE	0.1355
R-Square	0.3557
Adj R-Square	0.3387

Appendix B: Regression Assumptions SAT Mathematics



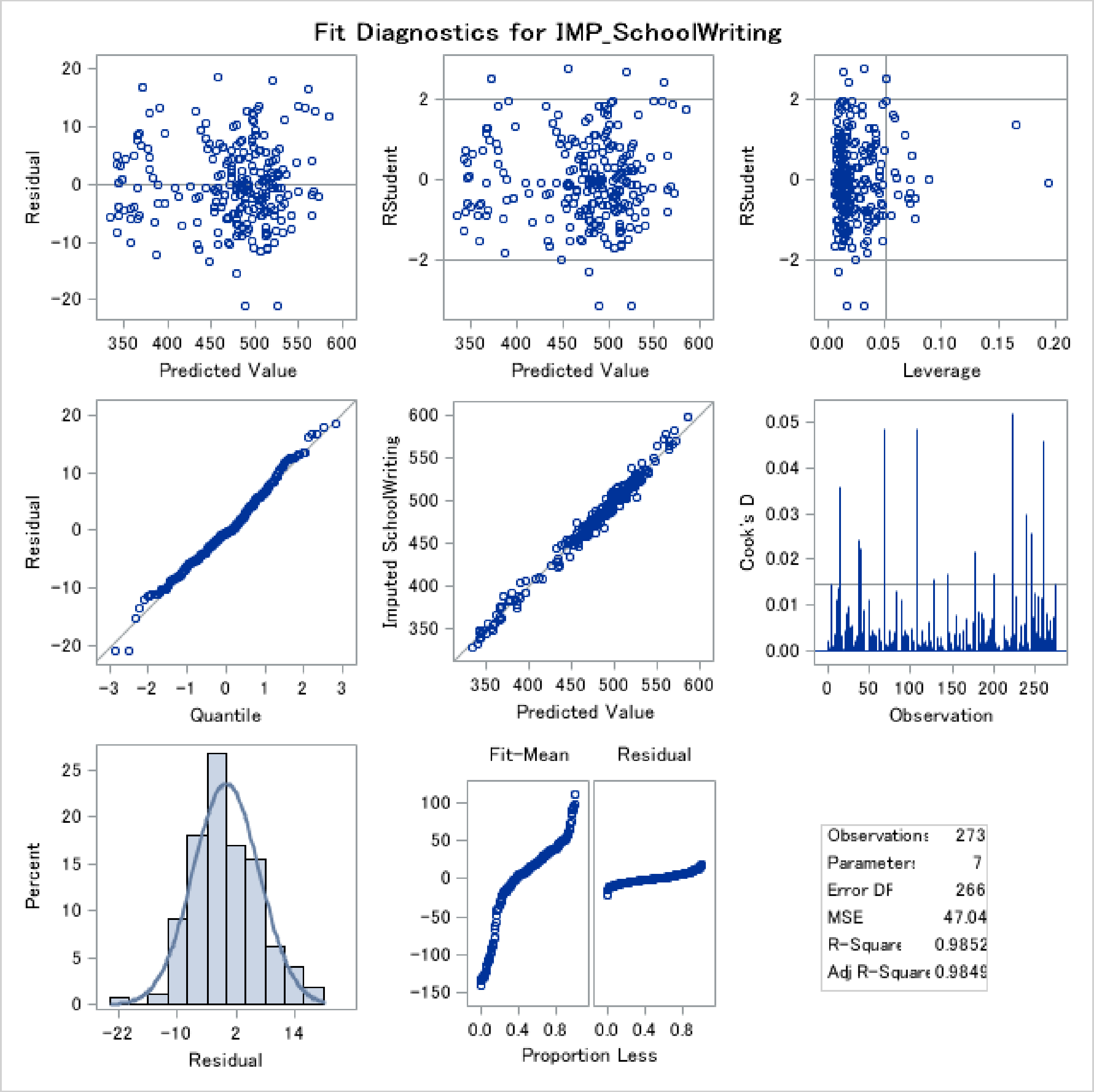
Observations	273
Parameters	10
Error DF	263
MSE	84.188
R-Square	0.9753
Adj R-Square	0.9745

Appendix C: Regression Assumptions SAT Reading

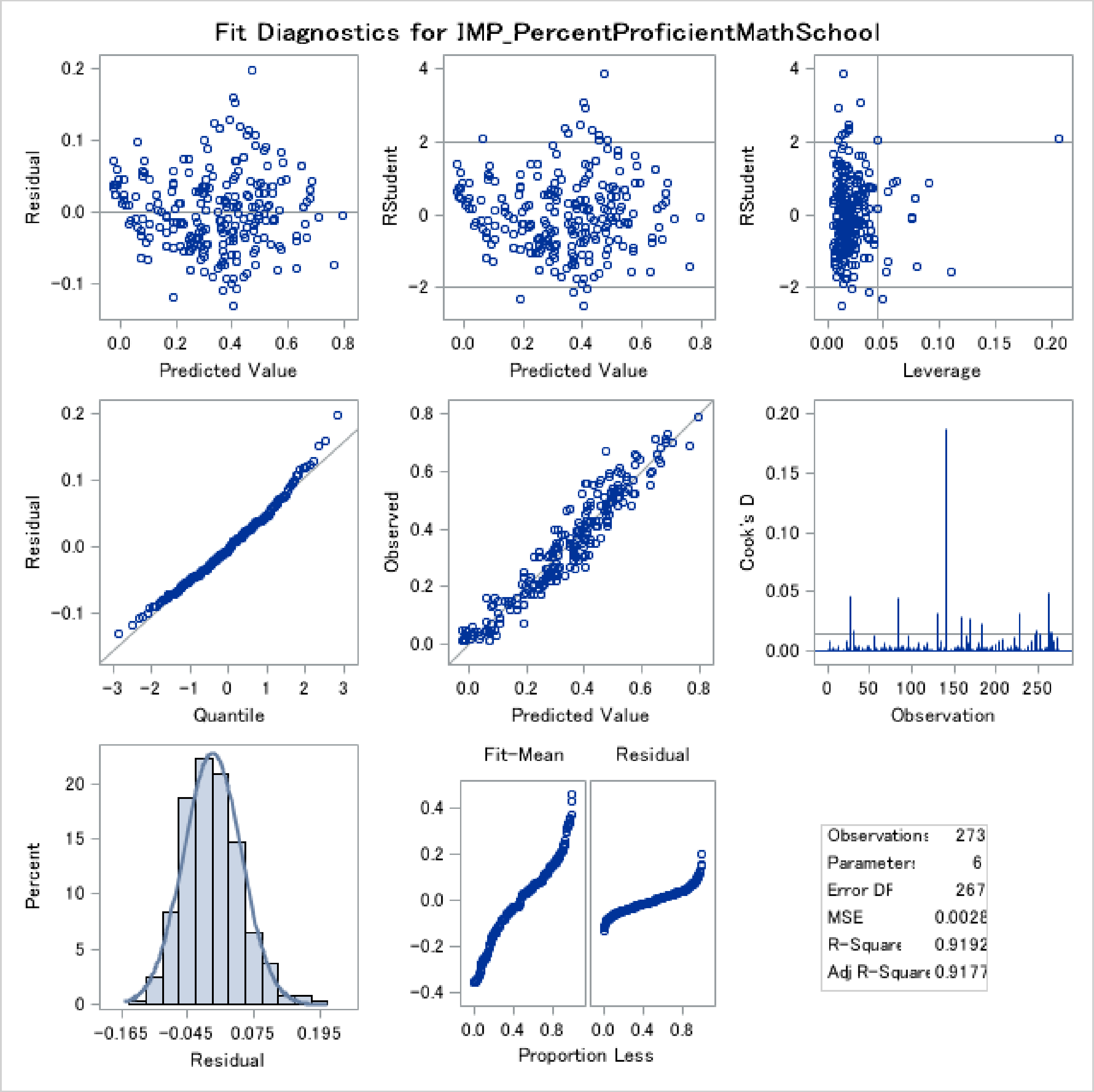


Observation	273
Parameter	6
Error DF	267
MSE	47.213
R-Square	0.9852
Adj R-Square	0.985

Appendix D: Regression Assumptions SAT Writing

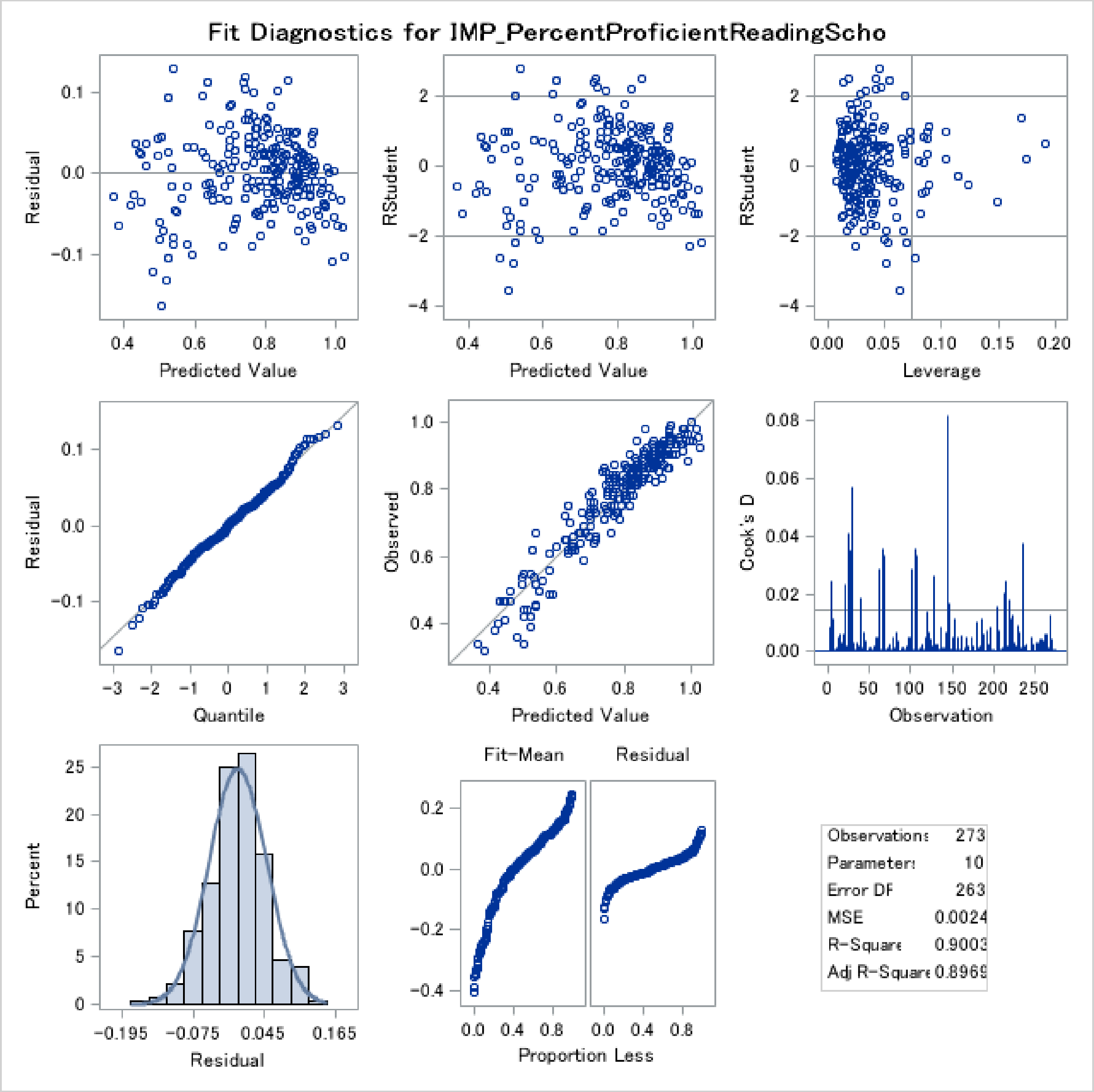


Appendix E: Regression Assumptions NECAP Math

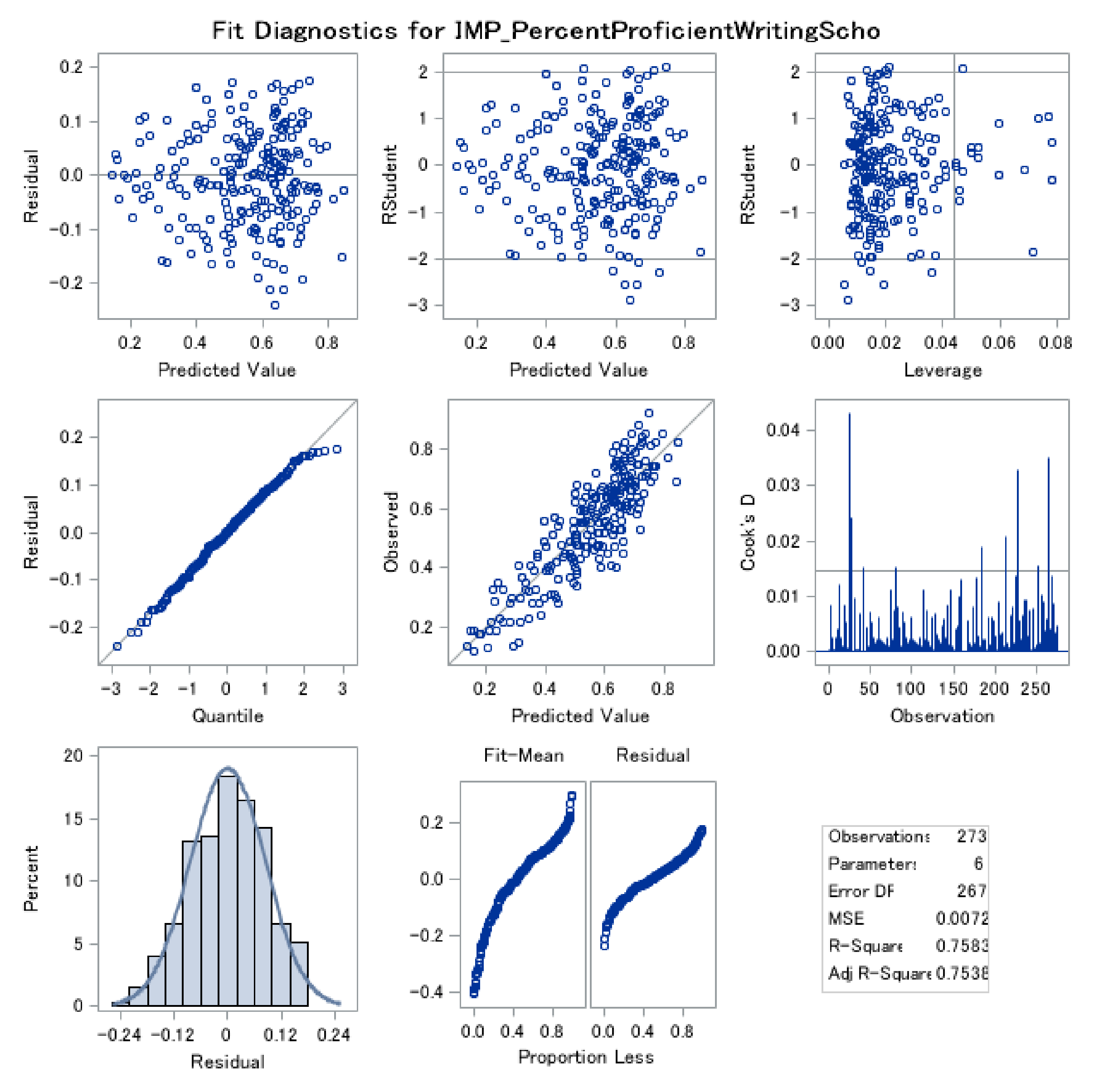


Observation:	273
Parameter:	6
Error DF	267
MSE	0.0028
R-Square	0.9192
Adj R-Square	0.9177

Appendix F: Regression Assumptions NECAP Reading

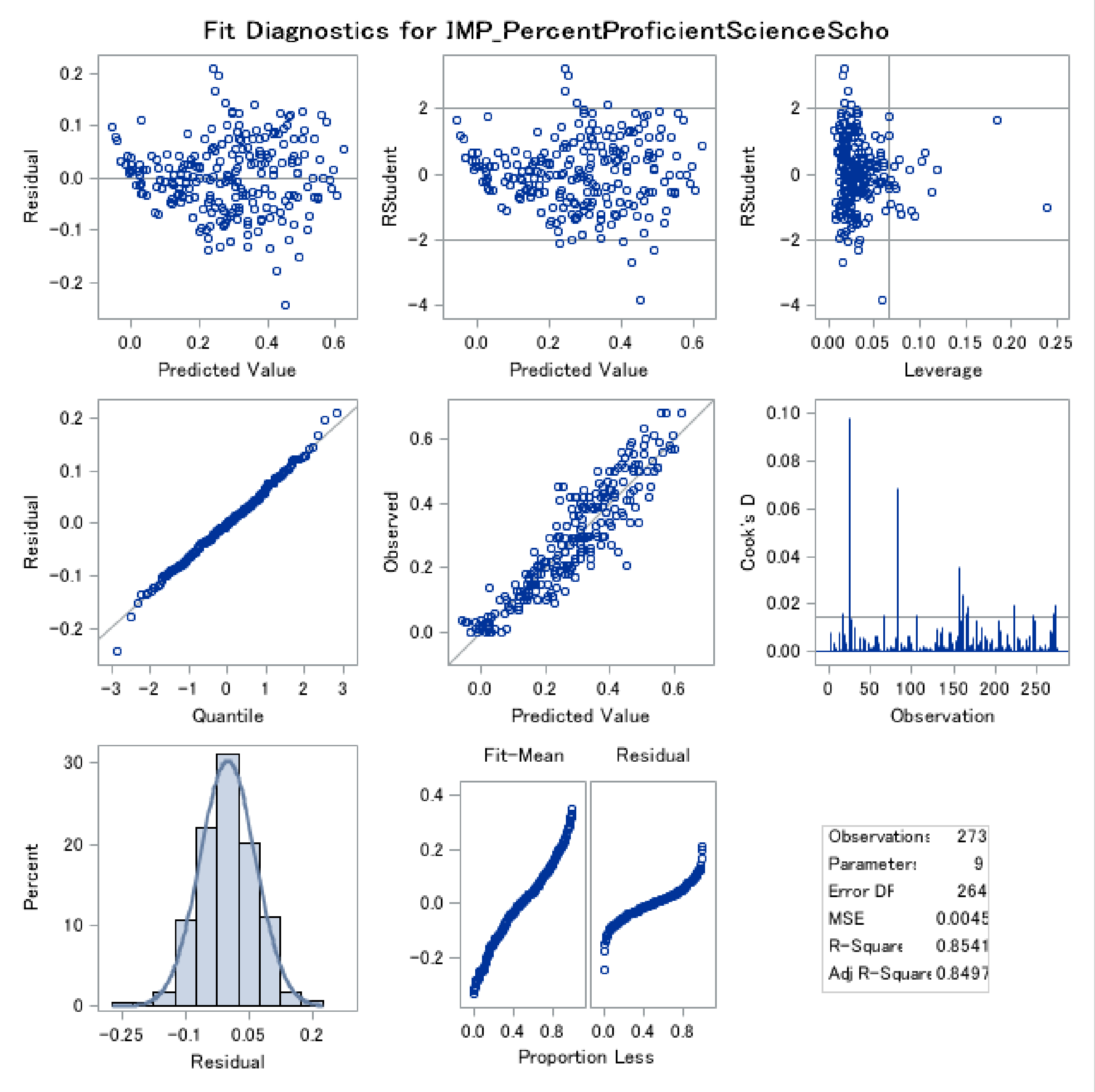


Appendix G: Regression Assumptions NECAP Writing



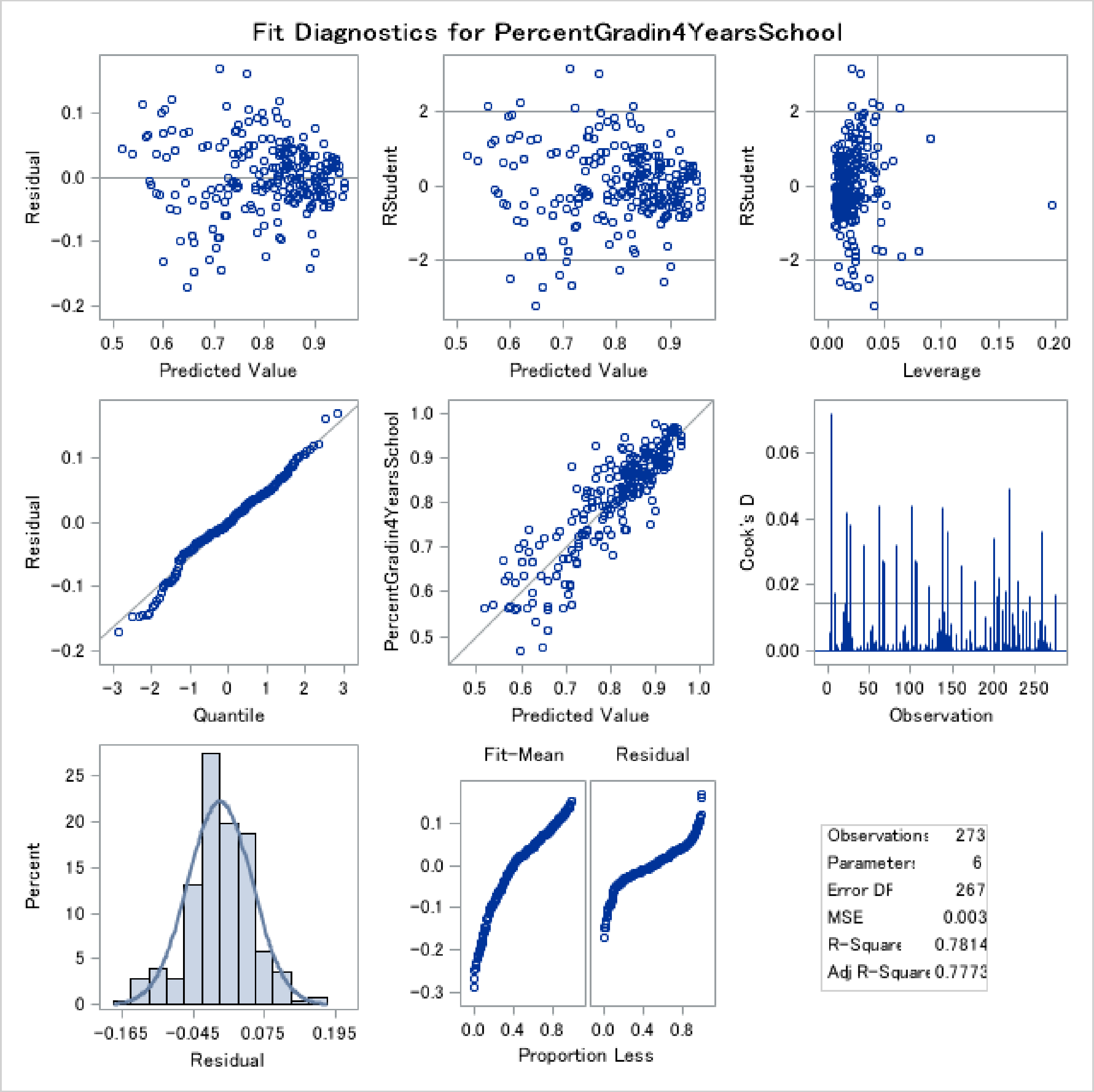
Observations	273
Parameter:	6
Error DF	267
MSE	0.0072
R-Square	0.7583
Adj R-Square	0.7538

Appendix H: Regression Assumptions for NECAP Science



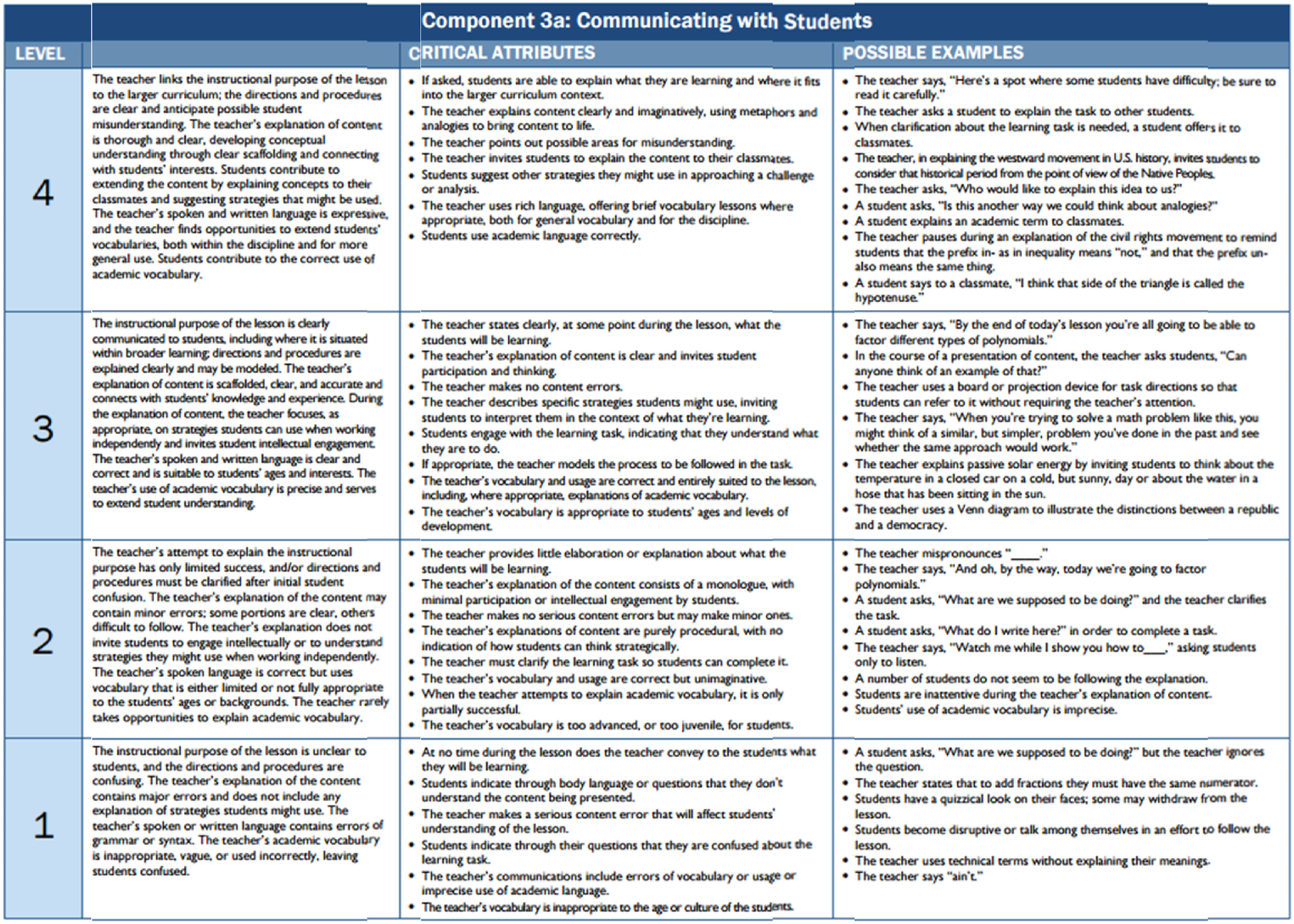
Observations	273
Parameters	9
Error DF	264
MSE	0.0045
R-Square	0.8541
Adj R-Square	0.8497

Appendix I: Regression Assumptions Four Year Graduation Rate



Observations	273
Parameters	6
Error DF	267
MSE	0.003
R-Square	0.7814
Adj R-Square	0.7773

Appendix J: Example Rubric for Teacher Evaluation (Component 3a)



	Component 3a: Communicating with Students		
LEVEL		CRITICAL ATTRIBUTES	POSSIBLE EXAMPLES
4	The teacher links the instructional purpose of the lesson to the larger curriculum; the directions and procedures are clear and anticipate possible student misunderstanding. The teacher's explanation of content is thorough and clear, developing conceptual understanding through clear scaffolding and connecting with students' interests. Students contribute to extending the content by explaining concepts to their classmates and suggesting strategies that might be used. The teacher's spoken and written language is expressive, and the teacher finds opportunities to extend students' vocabularies, both within the discipline and for more general use. Students contribute to the correct use of academic vocabulary.	• If asked, students are able to explain what they are learning and where it fits into the larger curriculum context. • The teacher explains content clearly and imaginatively, using metaphors and analogies to bring content to life. • The teacher points out possible areas for misunderstanding. • The teacher invites students to explain the content to their classmates. • Students suggest other strategies they might use in approaching a challenge or analysis. • The teacher uses rich language, offering brief vocabulary lessons where appropriate, both for general vocabulary and for the discipline. • Students use academic language correctly.	• The teacher says, "Here's a spot where some students have difficulty; be sure to read it carefully." • The teacher asks a student to explain the task to other students. • When clarification about the learning task is needed, a student offers it to classmates. • The teacher, in explaining the westward movement in U.S. history, invites students to consider that historical period from the point of view of the Native Peoples. • The teacher asks, "Who would like to explain this idea to us?" • A student asks, "Is this another way we could think about analogies?" • A student explains an academic term to classmates. • The teacher pauses during an explanation of the civil rights movement to remind students that the prefix in- as in inequality means "not," and that the prefix un- also means the same thing. • A student says to a classmate, "I think that side of the triangle is called the hypotenuse."
3	The instructional purpose of the lesson is clearly communicated to students, including where it is situated within broader learning; directions and procedures are explained clearly and may be modeled. The teacher's explanation of content is scaffolded, clear, and accurate and connects with students' knowledge and experience. During the explanation of content, the teacher focuses, as appropriate, on strategies students can use when working independently and invites student intellectual engagement. The teacher's spoken and written language is clear and correct and is suitable to students' ages and interests. The teacher's use of academic vocabulary is precise and serves to extend student understanding.	• The teacher states clearly, at some point during the lesson, what the students will be learning. • The teacher's explanation of content is clear and invites student participation and thinking. • The teacher makes no content errors. • The teacher describes specific strategies students might use, inviting students to interpret them in the context of what they're learning. • Students engage with the learning task, indicating that they understand what they are to do. • If appropriate, the teacher models the process to be followed in the task. • The teacher's vocabulary and usage are correct and entirely suited to the lesson, including, where appropriate, explanations of academic vocabulary. • The teacher's vocabulary is appropriate to students' ages and levels of development.	• The teacher says, "By the end of today's lesson you're all going to be able to factor different types of polynomials." • In the course of a presentation of content, the teacher asks students, "Can anyone think of an example of that?" • The teacher uses a board or projection device for task directions so that students can refer to it without requiring the teacher's attention. • The teacher says, "When you're trying to solve a math problem like this, you might think of a similar, but simpler, problem you've done in the past and see whether the same approach would work." • The teacher explains passive solar energy by inviting students to think about the temperature in a closed car on a cold, but sunny, day or about the water in a hose that has been sitting in the sun. • The teacher uses a Venn diagram to illustrate the distinctions between a republic and a democracy.
2	The teacher's attempt to explain the instructional purpose has only limited success, and/or directions and procedures must be clarified after initial student confusion. The teacher's explanation of the content may contain minor errors; some portions are clear, others difficult to follow. The teacher's explanation does not invite students to engage intellectually or to understand strategies they might use when working independently. The teacher's spoken language is correct but uses vocabulary that is either limited or not fully appropriate to the students' ages or backgrounds. The teacher rarely takes opportunities to explain academic vocabulary.	• The teacher provides little elaboration or explanation about what the students will be learning. • The teacher's explanation of the content consists of a monologue, with minimal participation or intellectual engagement by students. • The teacher makes no serious content errors but may make minor ones. • The teacher's explanations of content are purely procedural, with no indication of how students can think strategically. • The teacher must clarify the learning task so students can complete it. • The teacher's vocabulary and usage are correct but unimaginative. • When the teacher attempts to explain academic vocabulary, it is only partially successful. • The teacher's vocabulary is too advanced, or too juvenile, for students.	• The teacher mispronounces "____." • The teacher says, "And oh, by the way, today we're going to factor polynomials." • A student asks, "What are we supposed to be doing?" and the teacher clarifies the task. • A student asks, "What do I write here?" in order to complete a task. • The teacher says, "Watch me while I show you how to___," asking students only to listen. • A number of students do not seem to be following the explanation. • Students are inattentive during the teacher's explanation of content. • Students' use of academic vocabulary is imprecise.
1	The instructional purpose of the lesson is unclear to students, and the directions and procedures are confusing. The teacher's explanation of the content contains major errors and does not include any explanation of strategies students might use. The teacher's spoken or written language contains errors of grammar or syntax. The teacher's academic vocabulary is inappropriate, vague, or used incorrectly, leaving students confused.	• At no time during the lesson does the teacher convey to the students what they will be learning. • Students indicate through body language or questions that they don't understand the content being presented. • The teacher makes a serious content error that will affect students' understanding of the lesson. • Students indicate through their questions that they are confused about the learning task. • The teacher's communications include errors of vocabulary or usage or imprecise use of academic language. • The teacher's vocabulary is inappropriate to the age or culture of the students.	• A student asks, "What are we supposed to be doing?" but the teacher ignores the question. • The teacher states that to add fractions they must have the same numerator. • Students have a quizzical look on their faces; some may withdraw from the lesson. • Students become disruptive or talk among themselves in an effort to follow the lesson. • The teacher uses technical terms without explaining their meanings. • The teacher says "ain't."

Source: Rhode Island Department of Education

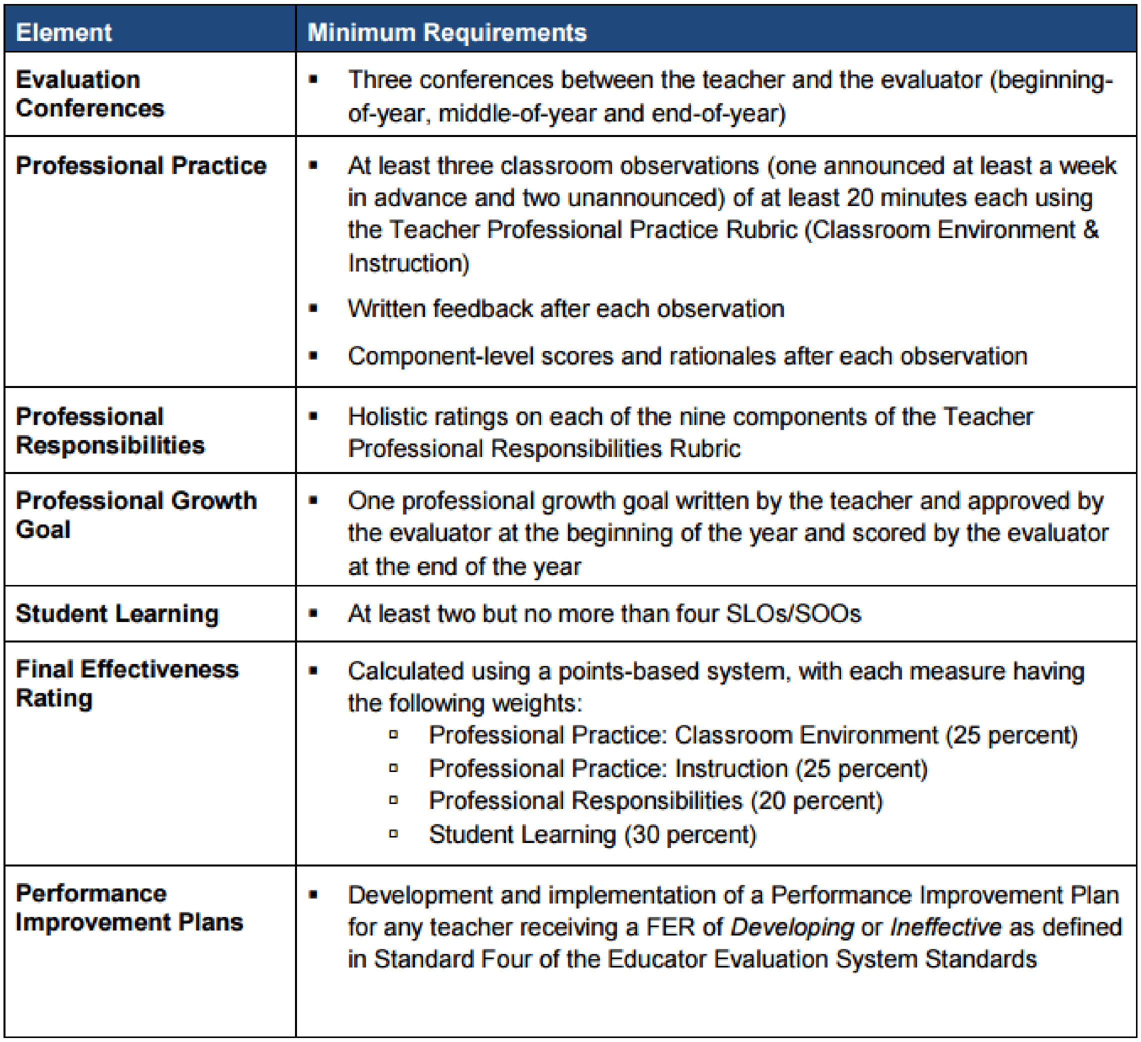
Appendix K: Variables

Dependent Variable	**Type**	**Description**
TeacherEval	Binary	1 if school year is at least 2012-2013, 0 otherwise
Four Year Grad Rate	Interval	0%-100%
SAT Math	Interval	200-800
SAT Reading	Interval	200-800
SAT Writing	Interval	200-800
NECAP Math	Interval	Percent Students Proficient
NECAP Reading	Interval	Percent Students Proficient
NECAP Writing	Interval	Percent Students Proficient
NECAP Science	Interval	Percent Students Proficient

Independent Variable	**Type**	**Description**
Attendance Rate	Interval	0%-100%
Chronic Absentee Rate	Interval	0%-100%
Classes Taught by Not Highly Qualified Teachers	Interval	Number of Classes
Eligibility for Subsidized Lunch	Interval	0%-100%
AP Exams Taken	Interval	Number of Exams Taken at School
Exams at College Mastery	Interval	Number of Exams with at least a 3
Drop Out Rate	Interval	0%-100%
GED Rate	Interval	0%-100%
Percent Teachers with Emergency Certification	Interval	0%-100%

Dependent Variable, continued	Type	Description
Alternate Placements	Interval	Number
In School Suspensions	Interval	Number
Out of School Suspensions	Interval	Number
Students taking AP Exams	Interval	Number
Students taking ESL	Interval	0%-100%
Students getting Special Education	Interval	0%-100%
Student-Teacher Ratio	Interval	Ratio

Appendix L: Teacher Evaluation Model in Rhode Island



Element	Minimum Requirements
Evaluation Conferences	▪ Three conferences between the teacher and the evaluator (beginning-of-year, middle-of-year and end-of-year)
Professional Practice	▪ At least three classroom observations (one announced at least a week in advance and two unannounced) of at least 20 minutes each using the Teacher Professional Practice Rubric (Classroom Environment & Instruction) ▪ Written feedback after each observation ▪ Component-level scores and rationales after each observation
Professional Responsibilities	▪ Holistic ratings on each of the nine components of the Teacher Professional Responsibilities Rubric
Professional Growth Goal	▪ One professional growth goal written by the teacher and approved by the evaluator at the beginning of the year and scored by the evaluator at the end of the year
Student Learning	▪ At least two but no more than four SLOs/SOOs
Final Effectiveness Rating	▪ Calculated using a points-based system, with each measure having the following weights: ▫ Professional Practice: Classroom Environment (25 percent) ▫ Professional Practice: Instruction (25 percent) ▫ Professional Responsibilities (20 percent) ▫ Student Learning (30 percent)
Performance Improvement Plans	▪ Development and implementation of a Performance Improvement Plan for any teacher receiving a FER of *Developing* or *Ineffective* as defined in Standard Four of the Educator Evaluation System Standards

Source: Rhode Island Department of Education

References

Balch, R., & Springer, M. G. (2015). Performance pay, test scores, and student learning objectives. *Economics Of Education Review*, *44,* 114-125. doi:10.1016/j.econedurev.2014.11.002.

Bengiamin, N. N., & Leimer, C. (2012). SLO-Based Grading Makes Assessment an Integral Part of Teaching. *Assessment Update*, *24*(5), 1-16.

Brimijoin, K. (2005). Differentiation and High-Stakes Testing: An Oxymoron? *Theory into Practice 44(3)*, 254-261. Retrieved from http://0-www.jstor.org.helin.uri.edu/stable/3497005.

Caillier, J. (2010). Paying teachers according to student achievement: Questions regarding pay-for-performance models in public education. *The Clearing House, 83*(2), 58-61. Retrieved from http://0-search.proquest.com.helin.uri.edu/docview/596621208?accountid=36823.

Gunzenhauser, M. (2003). High-Stakes Testing and the Default Philosophy of Education. *Theory into Practice 42(1)*, 51-58. Retrieved from http://0-www.jstor.org.helin.uri.edu/stable/1477318.

Hill, H., Kapitula, L., and Umland, K. (2011). A Validity Argument Approach to Evaluating Teacher Value-Added Scores. *American Educational Research Journal 48(3)*, 794-831. Retrieved from http://www.jstor.org/stable/27975308.

Holt, M. (2001). Performance pay for teachers: The standards movement's last stand? *Phi Delta Kappan, 83*(4), 312-317. Retrieved from http://0-search.proquest.com.helin.uri.edu/docview/218480405?accountid=36823.

Lazear, E. (2006). Speeding, Terrorism, and Teaching to the Test. *The Quarterly Journal of Economics 121(3)*, 1029-1061. Retrieved from http://www.jstor.org/stable/25098816

Liang, G., & Akiba, M. (2015). Teacher Evaluation, Performance-Related Pay, and Constructivist Instruction. *Educational Policy*, *29*(2), 375-401. doi:10.1177/0895904813492379.

Lynn, A. R. (2013). Teacher Evaluations Based on Student Testing: Missing an Opportunity for True Education Reform. *Texas Journal On Civil Liberties & Civil Rights*, *18*(2), 203-234.

Moore, S., & Kuol, N. (2005). Students evaluating teachers: exploring the importance of faculty reaction to feedback on teaching. *Teaching In Higher Education*, *10*(1), 57-73. doi:10.1080/1356251052000305534.

Podgursky, M. and Springer, M. (2007). The Single Best Idea to Improve K-12 Education. *Peabody Journal of Education 82(4),* 551-573. Retrieved from http://www.jstor.org/stable/25594760.

Rhode Island Department of Education. (n.d.). Retrieved from http://www.ride.ri.gov/.

Rockoff, J. & Speroni, C. (2010). Subjective and Objective Evaluations of Teacher Effectiveness. *The American Economic Review, 100(2),* 261-266. Retrieved from http://dx.doi.org/10.1257/aer.100.2.261.

Sass, T. R., Semykina, A., & Harris, D. N. (2014). Value-added models and the measurement of teacher productivity. *Economics Of Education Review*, *38*9-423. doi:10.1016/j.econedurev.2013.10.003.

Viadero, D. (2007). Study links merit pay to slightly higher student scores. *Education Week, 26*(18), 8. Retrieved from http://0-search.proquest.com.helin.uri.edu /docview/202753785?accountid=36823.