An Analysis of the Teacher Instructional Growth for Effectiveness and Results (TIGER) Model Pilot

Technical Report

Clint Sattler
Supervisor, Research and Evaluation
Knox County Schools
Department of Research, Evaluation, and Assessment

December 2018

Overview

The Teacher Instructional Growth of Effectiveness and Results (TIGER) model for teacher observation was piloted in eight Knox County Schools during the 2017-2018 school year (SY1718). The TIGER model is a state-approved alternative to the teacher observation system currently deployed in Knox County: the Tennessee Educator Acceleration Model (TEAM). Both models use a variation of the Charlotte Danielson Framework for Teaching to promote professional development through observation and feedback (Danielson Group). The TEAM and TIGER frameworks assess twenty-three indicators in four domains: Planning, Environment, Instruction, and Professionalism (TDOE 2016b). Both models include elements identified by the National Council on Teacher Quality as core requirements for effective evaluation systems (Putnam). However, the developers of the TIGER model claim their model is more focused on continuous reflection, provides more useful feedback to educators, and better fosters teacher improvement (The Teacher Instructional Growth for Effectiveness and Results).

The schools chosen to participate in the TIGER pilot were Farragut High, Halls Middle, Mount Olive Elementary, Powell High, Richard Yoakley, Spring Hill Elementary, West Haven Elementary, and West Valley Middle. These schools were accepted as the pilot schools because surveys conducted by Dr. Rodney Russel, former Director of Human Capital Strategy, indicated that 75% or more of the respondents at these schools were interested in exploring a new observation system.

This analysis seeks to answer to research questions related to the TIGER model pilot.

1. How did teacher perception of the evaluation system change after the TIGER model was adopted in the pilot schools?
2. What was the estimated impact of adopting the TIGER model on observation scores in the pilot schools?
3. What was the estimated impact of adopting the TIGER model on the relationship between observation score and teacher value-added (TVAAS) estimates in the pilot schools?

Readers should note that the findings from these analyses cannot be considered causal nor extrapolated to the district as a whole because of the deliberate selection bias in choosing the pilot schools.

Methodology: Teacher Perception Data

Data from the Tennessee Educators survey for the pilot schools were analyzed in order to detect changes in teacher perception of their evaluation system. There were three key questions that were used in the analysis.

1. The process used to conduct my teacher evaluation last year was fair to me.
2. In general, the teacher evaluation process used in my school last year led to improvements in my teaching.
3. In general, the teacher evaluation process used in my school last year led to improvements in student learning.

All questions were scored on a Likert-type scale (Strongly Disagree, Disagree, Agree, and Strongly Agree).

At a minimum, 45% of teachers at a school had to respond to the survey in order to generate data. Table 1 shows the schools that met the minimum participation rate by academic year.

Table 1: TN Educator Survey Participation

	Met Minimum TN Educator Survey Participation Requirements			
School	SY1718	SY1617	SY1516	SY1415
Farragut High	No	No	Yes	Yes
Powell High	No	No	No	Yes
Richard Yoakley	Yes	No	No	No
Halls Middle	No	No	No	Yes
West Valley Middle	No	No	No	Yes
West Haven Elementary	Yes	No	Yes	Yes
Mt. Olive Elementary	Yes	No	No	Yes
Spring Hill Elementary	No	Yes	No	Yes

The only year with enough data to serve as an adequate baseline was SY1415, which was three years prior to treatment. The SY1718 data collected by the state may not be representative of the perceptions at all of the pilot schools because the majority of schools did not meet the minimum participation requirement. District-level surveying (using the same three questions) occurred via Survey Monkey in the Fall of SY1819. The data collected during this subsequent survey was used to measure perception after treatment.

Methodology: Quantitative Analysis

Quantitative analysis required the creation of a control group of teachers using quasi-experimental methods. Coarsened exact matching (CEM) was chosen as the method for creating comparison groups. We used the k-to-k method of CEM to ensure the same number of teachers in the final treatment and control groups.

CEM requires the researcher to create cut points to "bin" continuous input variables. Bin sizes were set to maximize the number of matched teachers without biasing the results of statistical tests. All categorical variables required a direct match between treatment and control groups. Pay step, which was used as an estimate of years of experience, was binned at 1 year increments up to 10 years. All teachers with 10 or more years of experience could be matched to each other. This practice is supported by outside research (Papay).

Methodology: Impact on Observation Scores

Teachers in the treatment and control pools were paired by pay step, the percentage of students in their classes defined as economically disadvantaged, and their SY1617 observation score. Exploratory analysis indicated that the outcome variable (SY1718 observation score) was most closely correlated with the SY1617 observation score (Pearson R=0.76). The cut points for previous observation scores were set at 0.025 raw score points (5% of the standard deviation for this variable). Pay step and the percent economically disadvantaged exhibited less correlation with the outcome variable (Pearson R=0.25 and R=-0.26 respectively). Cut points for the percent economically disadvantaged were set at 5% (25% of the standard deviation). Cut points for pay step were set at 1 year increments for logical reasons.

Once k-to-k CEM matching created the treatment and control groups, an Anderson-Darling k-Samples Test was used to determine if the SY1718 observation scores of treatment and control groups were likely to have come from the same underlying distribution. Findings were considered significant at the $\alpha=0.05$ level. A 95% t-distribution confidence interval was created for the mean observation score for the control group. The estimated mean observation score for the treatment group was then compared to this confidence interval to determine if the mean was significantly different between the two groups.

Methodology: Relationship between TVAAS and Observation Scores

The analysis of the value-added data used the TVAAS index as the outcome variable. SAS (the vendor who computes TVAAS estimates) deploys different models to calculate teacher value-added estimates based on the grade-level taught. Value-added estimates for elementary and middle school teachers are derived from a multivariate model (MRM) that estimates teacher impact in normal curve equivalents. Value-added estimates for high school teachers are derived from a univariate model (URM) that estimates teacher impact in scaled score units. The TVAAS index is a non-dimensional number that can be compared

regardless of models (SAS). The sample sizes used for the analysis of TVAAS scores is considerably lower than the sample sizes used in the observation score analysis for research question #2. This is because only a subset of teachers in each pool have TVAAS indices in consecutive years because of the grade-levels and/or subjects they teach.

Teachers in the treatment and control pools were paired by input variables that had moderate to strong correlations with the SY1718 TVAAS index. These variables included SY1617 TVAAS index (Pearson R=0.44), SY1617 observation scores (Pearson R=0.22), the model used to calculate their index (MRM or URM), and the subject(s) taught. The calculation methodology (MRM or URM) and the subject(s) taught (Reading, Math, Science, Social Studies, or multiple subjects) were significant descriptors of variance in SY1718 TVAAS indices according to linear screening models (α=0.05).

A simple linear regression model of the SY1718 TVAAS index by SY1718 observation score and treatment condition was constructed after building a comparable pool of treatment and control teachers (Equation 1).

$$TVAAS\ Index_{SY1718} = Observation\ Score_{SY1718} + Treatment\ Condition + \varepsilon$$

Equation 1: TVAAS Regression Model

All calculations were done on R version 3.4.3 running on R Studio version 1.0.143. Coarsened exact matching was accomplished using version 1.1.19 of the cem package.

Results: Teacher Perception Data

The results of the pre-treatment and post-treatment surveys administered at the TIGER pilot schools can be found in Table 2. The data for the entire district (which would also include TIGER pilot school data) is contained in Table 3. Figure 1 allows for comparison between the two sets of survey respondents.

Table 2: TIGER Perception Data

	TIGER Post-Treatment: Fall SY1819 Survey, N=210				TIGER Pre-Treatment: 2015 TN Educator Survey, N=289			
Prompt	Strongly Disagree	Disagree	Agree	Strongly Agree	Strongly Disagree	Disagree	Agree	Strongly Agree
The process used to conduct my teacher evaluation last year was fair to me.	3%	5%	43%	48%	16%	33%	44%	7%
In general, the teacher evaluation process used in my school last year led to improvements in my teaching.	3%	15%	52%	30%	13%	30%	48%	10%
In general, the teacher evaluation process used in my school last year led to improvements in student learning.	5%	15%	50%	30%	15%	35%	41%	9%

Table 3: Entire District Perception Data

	All District Responses: 2018 TN Educator Survey, N=2,012				All District Responses: 2015 TN Educator Survey, N=2,766			
Prompt	Strongly Disagree	Disagree	Agree	Strongly Agree	Strongly Disagree	Disagree	Agree	Strongly Agree
The process used to conduct my teacher evaluation last year was fair to me.	7%	22%	57%	13%	12%	27%	51%	11%
In general, the teacher evaluation process used in my school last year led to improvements in my teaching.	7%	24%	56%	13%	9%	24%	53%	14%
In general, the teacher evaluation process used in my school last year led to improvements in student learning.	9%	29%	53%	10%	10%	30%	49%	11%

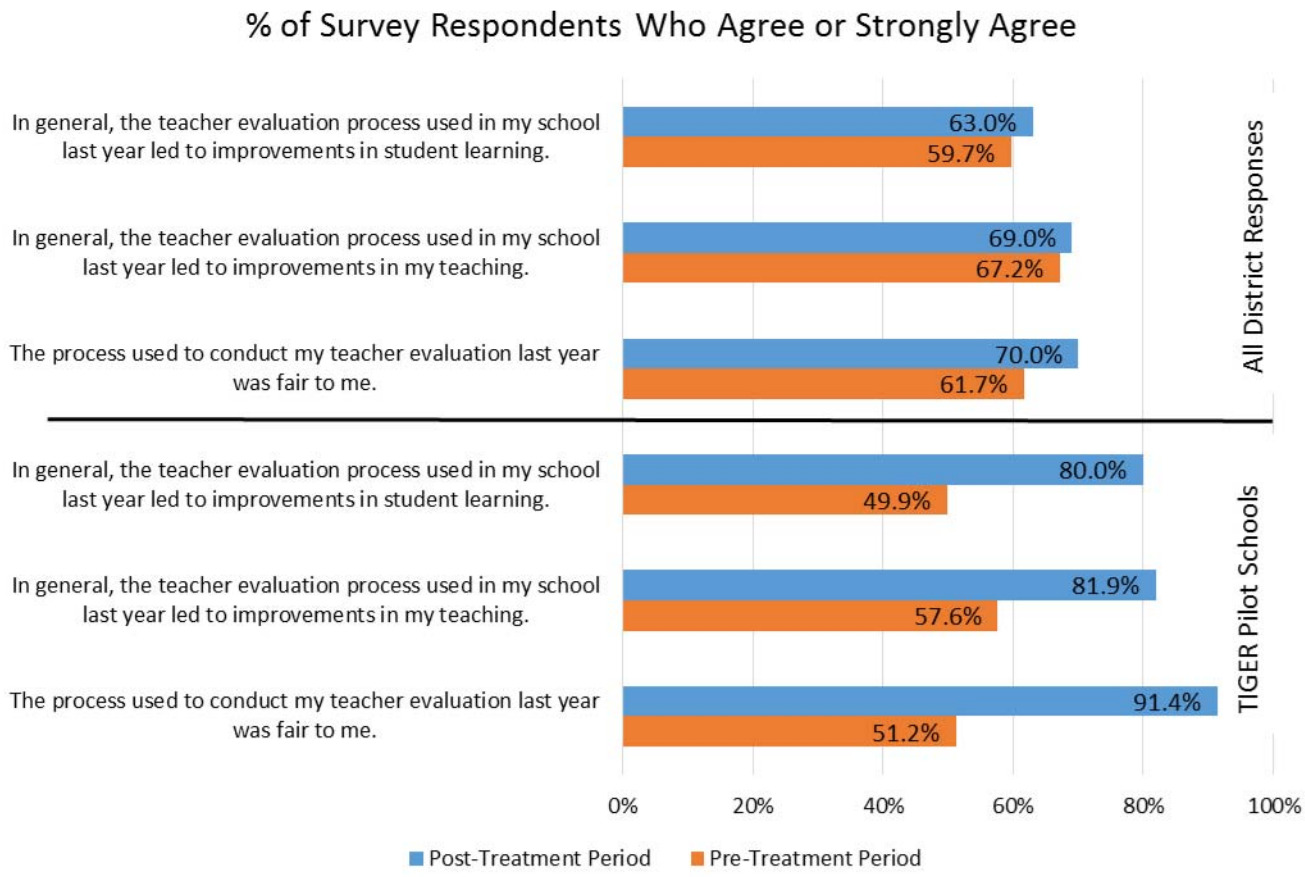


Figure 1: Comparison of Perception Data

Evidence suggests that the TIGER model positively impacted teachers' perceptions about the fairness of the evaluation system and impacts on teaching and learning at the pilot schools. The largest change in the percent of respondents agreeing or strongly agreeing was associated with the item, "the process used to conduct my teacher evaluation last year was fair to me."

The pre-treatment and post-treatment distributions of TVAAS indices at TIGER schools do not reflect teacher perception. The distribution of teacher TVAAS indices shifted downward in SY1718 when compared to SY1617 despite a high percentage of teachers agreeing or strongly agreeing that the TIGER evaluation system led to improvements in teaching and learning (Figure 2). It is possible that the TVAAS index was not aligned with the specific increases in teaching and learning perceived by the survey respondents or that TVAAS may not be very sensitive to changes in teacher practices (Chetty).

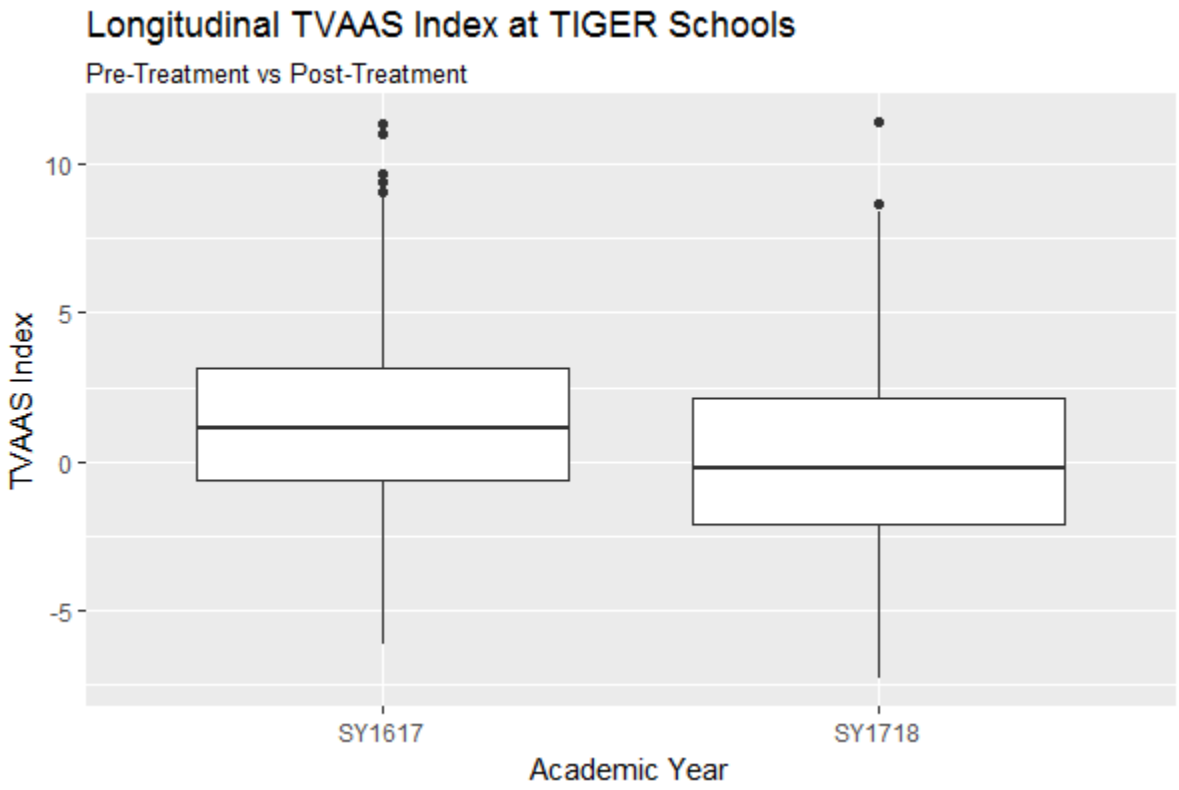


Figure 2: Distribution of TVAAS Indices

Results: Impact on Observation Scores

The results of the k-to-k CEM matching is contained in Table 4. The CEM algorithm was able to match 54.9% of teachers in the TIGER pilot with teachers observed under the TEAM system.

Table 4: CEM Matching Statistics: Observation Scores

Status	Control (TEAM)	Treatment (TIGER)
All	2757	350
Matched	192	192
Unmatched	2565	158
% Matched	7.0%	54.9%

In order to match, the teachers had to share similar characteristics in pay step, SY1617 observation score, and the percent of their students who were classified as economically disadvantaged. We achieved a multivariate imbalance measure L1 = 0.00. A density histogram of the SY1617 observation score for the matched samples is shown in Figure 3 to illustrate the level of matching.

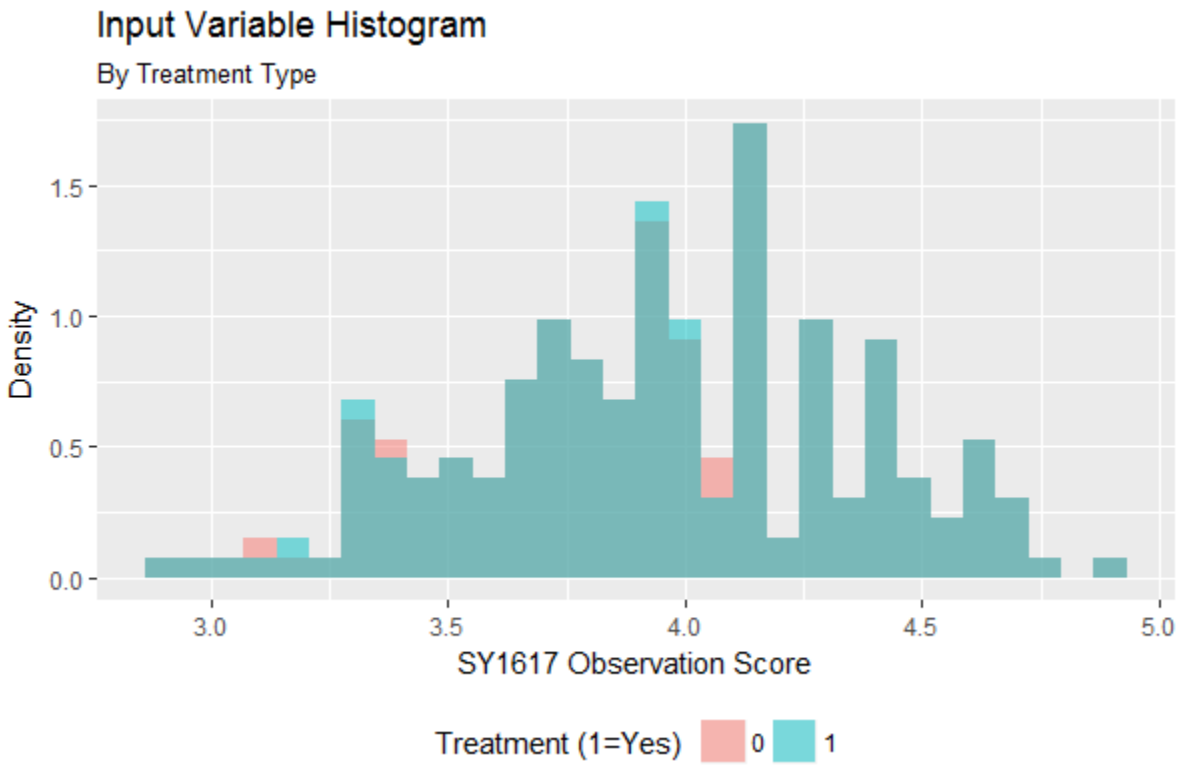


Figure 3: Sample Input Variable Density Histograms, Observation Scores

The results of the Anderson-Darling test indicate that we can reject the null hypothesis that the SY1718 observation scores were derived from a common population (AD statistic=5.055, 1000 simulations, p-value=0.0026). The shift in the outcome variable is evident in the density histogram in Figure 4. The distribution of SY1718 observation scores for teachers observed under the TIGER model was skewed towards higher scores when compared to the TEAM score distribution.

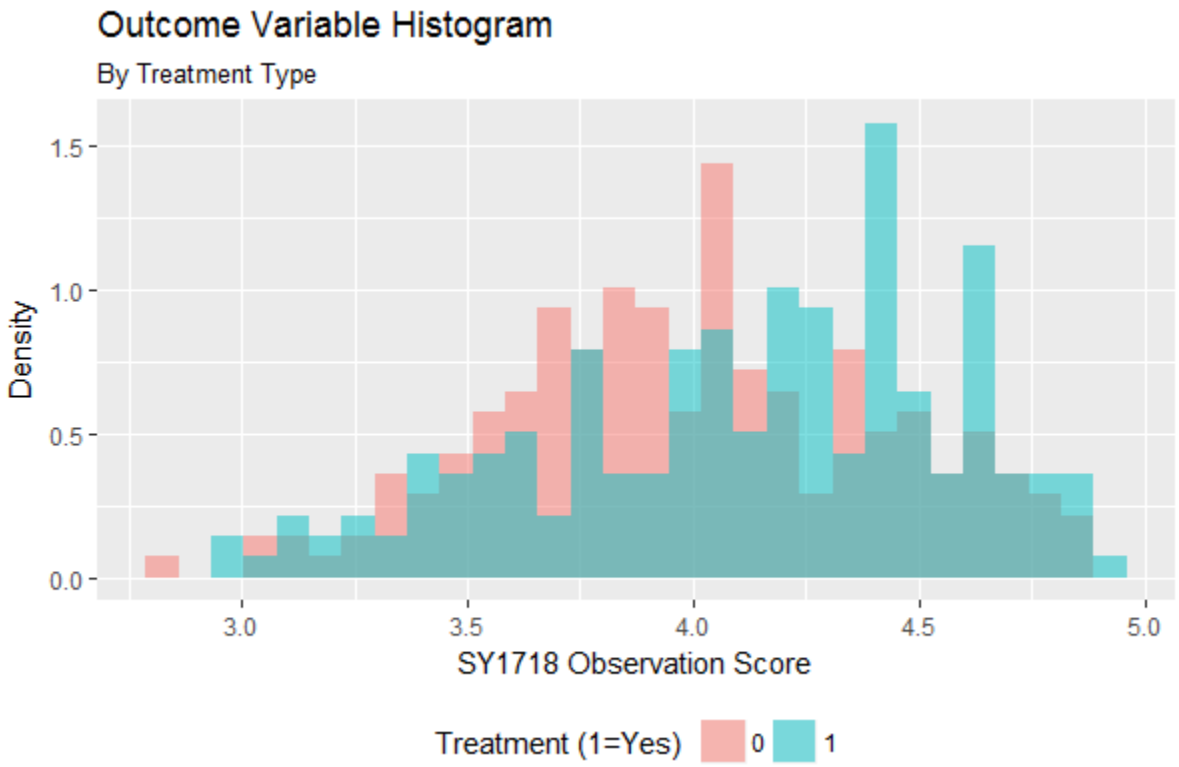


Figure 4: Outcome Variable Density Histograms

The 95% confidence interval for the mean SY1718 observation score for the control group was 3.96 to 4.08. The mean SY1718 observation score for the treatment group was 4.10, which is greater than the upper confidence interval for the control group. There is statistically significant evidence that the mean SY1718 observation score at TIGER schools was greater than the mean observation score at TEAM schools.

Results: Relationship between TVAAS and Observation Scores

The results of the k-to-k CEM matching is contained in Table 5. The CEM algorithm was able to match 77.7% of teachers in the TIGER pilot with teachers observed under the TEAM system.

Table 5: CEM Matching Statistics: Observation Scores

Status	Control (TEAM)	Treatment (TIGER)
All	732	112
Matched	87	87
Unmatched	645	25
% Matched	11.9%	77.7%

In order to match, the teachers had to share similar characteristics in the TVAAS model used in their calculation, SY1617 TVAAS index, SY1617 observation score, and the subject(s) in which their TVAAS index was generated. We achieved a multivariate imbalance measure L1 = 0.380. The greatest imbalance was in the matching of SY1617 TEAM observation score (univariate L1=0.12). The imbalance associated with the SY1617 TVAAS index was much less (univariate L1=0.02). This was expected. The CEM algorithm was calibrated to provide a better match on SY1617 TVAAS index because of its stronger correlation with the outcome variable.

In order to answer the pertinent research question, we must ensure that the matching created groups of teachers for whom the relationship between TVAAS index and observation score was no different prior to the deployment of the TIGER pilot. Linear regressions of the relationship between SY1617 TEAM score and SY1617 TVAAS index are provided in Figure 5. The level of overlap in the regression lines provides strong evidence that the relationship between TVAAS index and observation score was no different among the (eventual) TIGER pilot teachers and their TEAM counterparts prior to treatment.

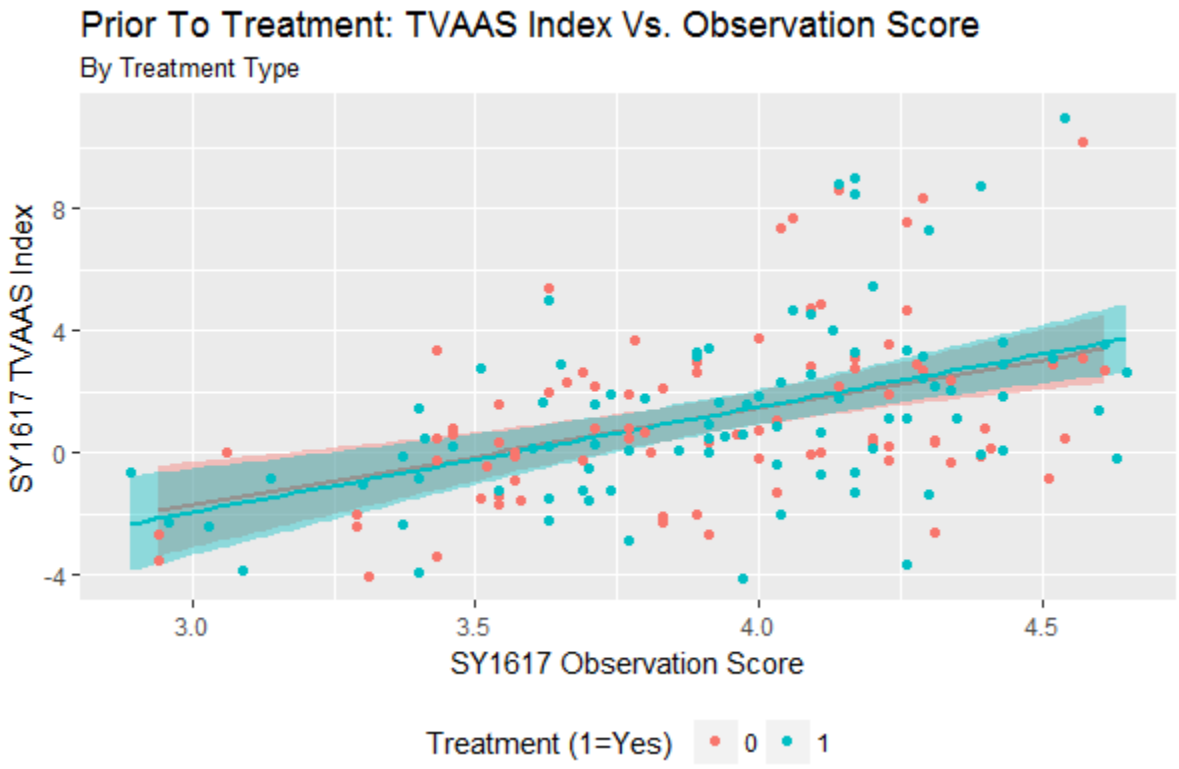


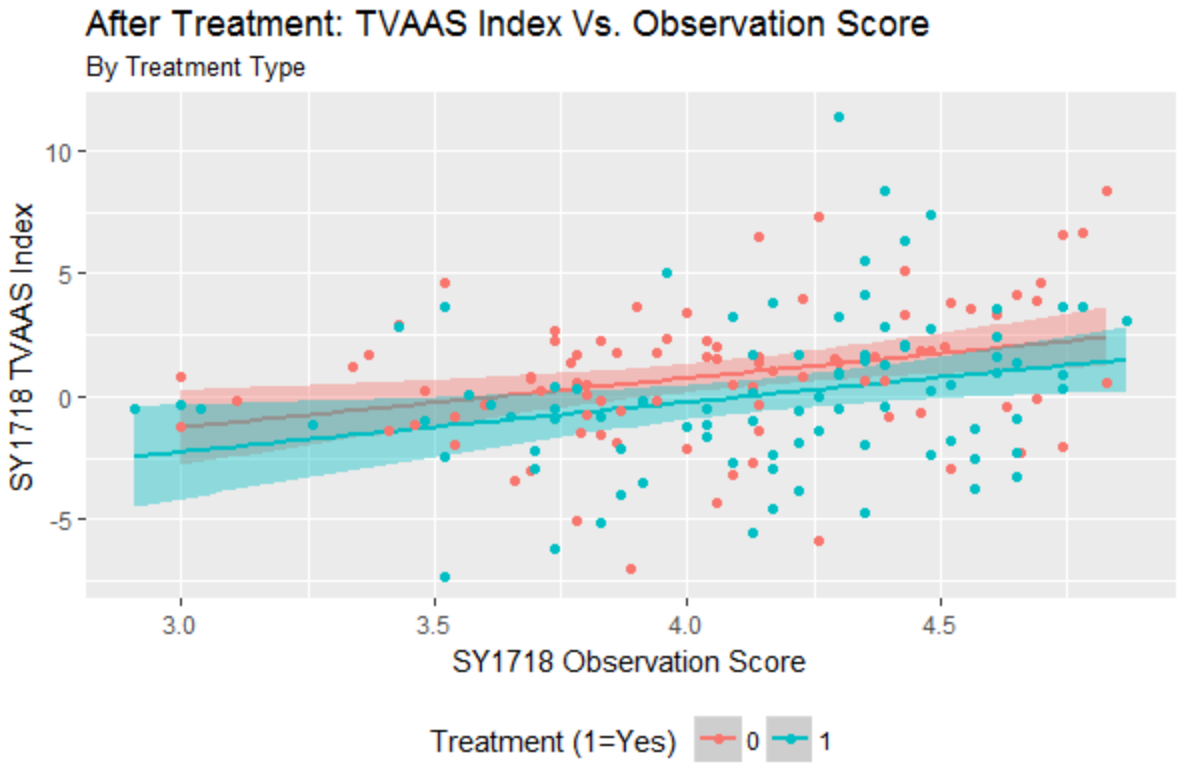
Figure 5: Prior to Treatment Relationship

The model parameters from the linear regression are contained in Table 6. The results indicate that we can reject the null hypothesis that there is no relationship between treatment condition and the linear relationship between SY1718 TVAAS index and SY1718 observation score.

Table 6: Linear Regression Output: Dependent Variable=SY1718 TVAAS Index

Parameter	Estimate	Std. Error	t value	p
Intercept	-7.33	2.11	-3.47	0.001
SY1718 Observation Score	2.02	0.52	3.91	0.000
Treatment	-0.98	0.44	-2.22	0.028

The relationship between SY1718 TVAAS index and SY1718 observation score can be seen in Figure 6. The results suggest that, at equivalent observation scores, teachers observed under the TIGER pilot had TVAAS indices that were approximately 1 point less than their TEAM-observed counterparts. The difference appears to be uniform across the entire distribution of observation scores.



Conclusions & Considerations

The analyses conducted by the Department of Research, Evaluation, and Assessment allows us to answer the three research questions related to the TIGER pilot conducted during SY1718 in the Knox County Schools.

How did teacher perception of the evaluation system change after the TIGER model was adopted in the pilot schools?

Evidence strongly suggested that teacher perception of the fairness of the evaluation system and its impact on teaching and learning were all positively impacted by the change to the TIGER model. An analysis of near-term TVAAS results suggests that any increases in instructional quality or student learning did not manifest itself on the state test. Increases in observation scores may indicate that teaching practice improved under the TIGER model, or it may indicate that the TIGER model is too different from TEAM for the direct comparison of scores collected under these competing models.

What was the estimated impact of adopting the TIGER model on observation scores in the pilot schools?

There is evidence to suggest that moving to the TIGER model corresponded to an increase in teacher observation scores. It is possible that instructional practices were positively impacted by the change to TIGER, as teacher responses to survey items suggest. However, there is little near-term outcome data to corroborate this. It is also possible that the data collection process for TIGER led to increases in observation scores. Observation scores in the TIGER model are based on a continuous yearlong process. In the TEAM process, indicators are scored based on the observations that occur during a discrete lesson. It is possible that the different data collection processes could yield different scores for the same teachers.

What was the estimated impact of adopting the TIGER model on the relationship between observation score and teacher value-added (TVAAS) estimates in the pilot schools?

Evidence suggests that the relationship between observation score and TVAAS index was different at the TIGER schools when compared to TEAM schools. Teachers at TIGER pilot schools were likely to have significantly lower TVAAS indices than a TEAM teacher with the same observation score (after controlling for background characteristics). The magnitude of this difference appears to be consistent across the entire distribution of observation scores.

The results of this analysis suggests that even though the TEAM and TIGER models use the same rubric, there are enough differences in implementation that the results from both models may not be equivalent. TEAM and TIGER observation scores may not be directly comparable.

It does not seem prudent to classify either system as better than the other. According to the Tennessee Department of Education, a successful evaluation system should "foster continuous improvement and innovation" (TDOE 2016a). There may be schools in which one observation model would meet this goal more readily than the other. The TIGER model may be the preferred evaluation system in settings in which teachers are not engaging with the current evaluation process because of perceived unfairness in the system. TEAM would likely be the preferred model in environments where the possibility of observation score inflation would be a barrier to reaching long-term strategic goals. Ideally, the state and the district would allow for the deployment of either model depending on the strategic priority of the school.

References

Chetty, R., Friedman, J. Rockoff, J. (2014). *Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates*. American Economic Review, 104(9), 2593-2632.

Danielson Group. (2017). *The Framework*. Retrieved from http://www.danielsongroup.org/framework/

Papay, J, Laski, M. (2018). *Exploring Teacher Improvement in Tennessee: A Brief Reimagining State Support for Professional Learning*. TN Education Research Alliance.

Putnam, H., Ross, E., Walsh, K. (2018). *Making a Difference: Six Places Where Teacher Evaluation Systems are Getting Results*. National Council on Teacher Quality (NCTQ).

SAS. (2017). *Technical Documentation of 2017 TVAAS Analyses*. SAS Institute.

The Teacher Instructional Growth for Effectiveness and Results (TIGER). (2018). *TIGER: A Formative Model for Teacher Evaluation*. Retrieved from https://tigermodel.net/about/

Tennessee Department of Education (TDOE). (2016a). *Teacher and Administrator Evaluation in Tennessee: A Report on Year 4 Implementation*.

Tennessee Department of Education (TDOE). (2016b). *Tennessee Educator Acceleration Model (TEAM) Teacher Evaluation System Handbook*. National Institute for Excellence in Teaching.