The Correlation between TEAM Observation Data and Current and Lagged TVAAS Index

Technical Report

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

March 2020

Overview

The Tennessee Educator Acceleration Model (TEAM) was officially launched during the 2011-2012 school year (SY1112) in the Knox County Schools (KCS). TEAM combines information collected on a standardized observation rubric with teacher-level and/or school-level performance data to derive a quantitative evaluation score. These scores may be used to inform high-stakes human capital decisions (such as teacher retention and/or promotion), teaching license advancement, and tenure determinations.

The implementation of TEAM (or an equivalent teacher evaluation system) was a legislative requirement following the state of Tennessee's winning of the federal Race-to-the-Top grant. The Tennessee Department of Education (TDOE) therefore had a vested interest in closely monitoring the efficacy of the TEAM process. The TDOE Division of Research and Analysis annually monitors key data related to TEAM efficacy. One metric used to monitor the TEAM process is the correlation between rubric scores from classroom observations and teacher growth scores (from the Tennessee Valued-Added Assessment System: TVAAS) which are estimated from state test data.

The emphasis that TDOE and KCS leadership placed on the alignment between TEAM observation scores and teacher TVAAS scores has varied through the years. Anecdotally, some school leaders experience implied pressure to align TEAM observation scores with TVAAS measures. Based on these comments, the KCS department of Research, Evaluation, and Assessment (REA) commissioned a study to determine how TEAM observation scores and teacher TVAAS scores related to each other in two consecutive school years. Our primary research question is, "Are current-year observation scores better correlated with current-year TVAAS estimates or previous-year TVAAS estimates?" The results of this analysis indicate that teachers' SY1819 observation scores correlated more strongly with previous-year TVAAS data. There may be policy implications from this finding.

Methodology

More than 4,000 teachers were evaluated with the TEAM model in SY1819. However, a majority of these teachers provided instruction in more than one content area. This threatens the validity of a study regarding the alignment between teacher TVAAS and classroom observation scores because the observation scores may not stem from the same content area in which TVAAS scores were generated. The export of the TEAM scores from the state database does not indicate the content area in which the observation data was collected, so there was no way to link TEAM data to the content area of the observed lesson. The sample of teachers used in this analysis was accordingly restricted to only include teachers who provided instruction in a single content area in which a state test is administered.

Teacher schedule data was extracted from the Knox County Student Information System (ASPEN) for the two most recently completed academic years (SY1718 and SY1819). Teachers were removed from the analysis if they provided instruction in multiple content areas in SY1718 and/or SY1819. Teachers were also excluded from the analysis if they did not provide instruction in the same content area in both years. Teachers that were observed under a pilot of the Teacher Instructional Growth for Effectiveness and Results (TIGER) model in SY1718 or SY1819 were likewise excluded.

The analysis was further restricted to only include teachers in grades 4-8. Technical manuals for the TVAAS calculation suggest that the growth estimates in these grades are more precise due to the statistical model used to estimate growth. Teachers with science and social studies data could not be included in this analysis due to longitudinal interruptions in operational state tests in these content areas. The TVAAS index (a teacher's growth estimate divided by the standard error of the growth estimate) was chosen as the input variable. The TVAAS index is directly used to derive a (categorical) TVAAS performance level. This TVAAS level is the metric referenced in TDOE TEAM-TVAAS alignment studies.

The relationship between SY1819 TEAM observation scores, SY1718 (single year) TVAAS indices and SY1819 (single year) TVAAS indices was determined using linear regression. A course content variable was included in case observation scores were influenced by the content area in which instruction was delivered. The final regression equation is as follows:

$$TEAM\ Score_{SY1819,i} = \beta_0 + \beta_1 * Content_i + \beta_2 * TVAAS_{SY1718,i} + \beta_3 * TVAAS_{SY1819,i}$$

Where TEAM $Score_{SY1819,i}$ is the geometric mean of all TEAM rubric scores for teacher i in SY1819, $Content_i$ is the content area (math or Language Arts) in which teacher i provided instruction in both SY1718 and SY1819, $TVAAS_{SY1718,i}$ is the TVAAS index associated with teacher i in SY1718 in $Content_i$, and $TVAAS_{SY1819,i}$ is the TVAAS index associated with teacher i in SY1819 in $Content_i$. This study included a total of 259 teachers (135 Language Arts teachers and 124 math teachers).

The results of the regression may be biased because the SY1718 TVAAS index and the SY1819 TVAAS index should not be independent of each other. Screening analysis indicates that the SY1718 TVAAS index is not highly collinear with the SY1819 TVAAS index. However, the Durbin Watson statistic associated with the TVAAS data is 1.74. The p value associated with this statistic is p=0.42, which indicates some level of auto-correlation between the TVAAS data. Some caution is recommended when interpreting the results of the regression analysis. Pair plots with the Pearson correlation coefficients of each variable with the SY1819 observation score are also provided.

Teacher observation data was extracted from the TNCompass platform (version 1.1). Teacher TVAAS data was extracted from TDOE accountability files in each respective year. Linear regression used R version 3.6.1 running on RStudio version 1.2.1335. Pair plots were generated using the psych package version 1.9.12.31. Regression diagnostics were monitored using the car package version 3.0-7.

Results

The statistics associated with the data used in the study are available in Table 1.

Table 1: Observation and TVAAS Data Summary

Measure	Min.	1st Qtile	Median	3rd Qtile	Max.	Mean	St. Dev
SY1819 Observation Score	2.63	3.78	4.09	4.36	4.96	4.06	0.42
SY1718 TVAAS Index	-7.31	-1.85	-0.15	1.28	7.35	-0.15	2.52
SY1819 TVAAS Index	-10.41	-1.47	0.02	1.61	13.23	0.03	2.80

The regression model provided a significant fit to the data (F=14,74, DoF=255, p-value=7.21e-9). The multiple R^2 for the fit was 0.147 with an adjusted R^2 of 0.137. The relatively low R^2 values suggest that there are likely other significant sources of variation in observation scores that are independent of the information contained in the TVAAS index. Diagnostics indicate that the residuals are normally distributed and homoscedastic. As mentioned in the methodology section, the results of the Durbin Watson test indicates that the results of the regression may be biased by non-independence of error terms.

The parameter estimates, 95% confidence intervals for the parameter estimates, and associated parameter statistics from the regression are contained in Table 2.

Table 2: Linear Regression Parameter Statistics

					Confidence Interval	
Parameter	Estimate	Std. Error	t value	Pr(>|t|)	2.5%	97.5%
Intercept	4.093	0.035	118.05	< 2e-16	4.02	4.16
Subject: Math	-0.052	0.051	-1.023	0.31	-0.15	0.05
SY1718 TVAAS Index	0.056	0.010	5.39	1.6e-7	0.04	0.08
SY1819 TVAAS Index	0.016	0.010	1.71	0.089	-0.002	0.04

Both of the TVAAS indices could be considered significant (or nearly significant) predictors of SY1819 observation scores, but the effect size of the SY1718 TVAAS index is larger than that of SY1819. Additionally, the 95% confidence interval provides more evidence that the SY1718 TVAAS index is more significantly correlated with the SY1819 observation score. The results indicate the content area in which the data was generated had no significant impact on the SY1819 observation score and was not necessary to include in the model.

The pair plot for the SY1819 observation score, SY1718 TVAAS index, and SY1819 TVAAS index is contained in Figure 1. Bi-variate Pearson correlation coefficients (R) are reported with the corresponding histograms and scatter plots.

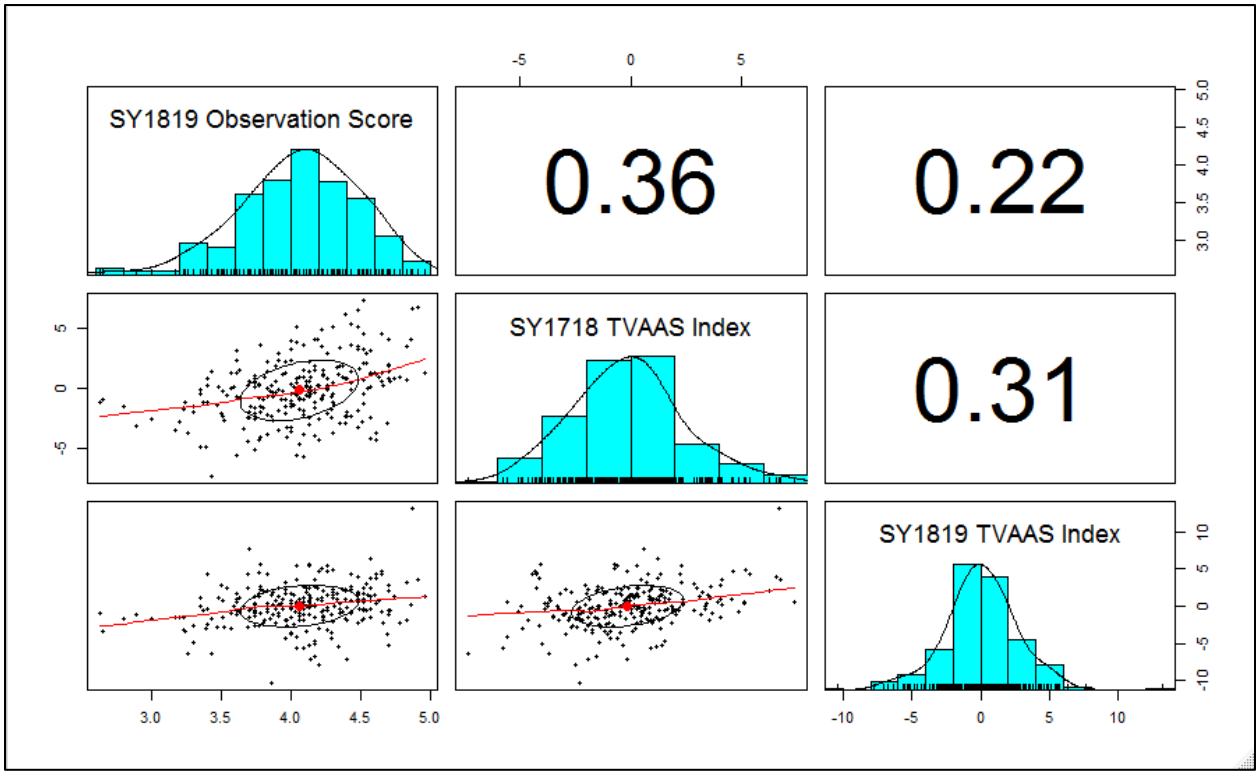

Figure 1: Observation Score and TVAAS Index Pair Plot

The pair plot indicates that the Pearson correlation coefficient is larger for the SY1819 observation score-SY1718 TVAAS index pairing than the SY1819 observation score–SY1819 TVAAS index pairing. The SY1718 TVAAS index describes approximately 8.1% more variation in the SY1819 observation score than the SY1819 TVAAS index. Interestingly, the SY1819 observation score-SY1718 TVAAS index pairing has a larger Pearson correlation coefficient than the SY1718 TVAAS index-SY1819 TVAAS index pairing.

Conclusions & Considerations

The data indicates that teachers' SY1819 observation scores correlate more strongly with their previous-year TVAAS data when compared to the current-year TVAAS data. This finding may be surprising since the SY1819 TVAAS data was generated from the instructional activities that were scored on the SY1819 TEAM observation rubrics.

It is possible that this finding is a result of random chance. The TEAM observation data is a criterion-based measure collected during discreet classroom observations. The scoring criteria for TEAM components are defined by the standards of the TEAM rubric. The TVAAS measure, however, is an aggregate measure of relative growth on the state assessment. It is possible that random chance results in the lagged TVAAS index correlating more strongly with current-year observation scores. However, if this alignment does arise by chance, these findings may provide evidence that monitoring the relationship between norm-referenced growth and criterion-based observation scores to estimate observation efficacy may be poor policy. Unfortunately, the methodology cannot be replicated using data from other years because of changes in the TVAAS calculation methodology and missing state test data. An inability to reproduce these findings (if such data were available) would better indicate that chance dictates the strength of the correlation between observation scores and TVAAS indices.

It is also possible that policy places pressure on observers to align observation scores with known TVAAS scores. Observers have no way of knowing what the current-year TVAAS scores will be for a given teacher. It seems plausible that if an observer felt pressure to align observation scores with TVAAS metrics then they would calibrate their observation scores to the available (lagged) TVAAS measure. Policy makers within the state and the district may wish to re-evaluate their policy regarding alignment of these measures to ensure that these policies are having their intended effects upon the teacher evaluation system.