Teacher Performance Evaluations and Value-Added Scores: Evidence from North Carolina Public Schools

By

Dayne Batten

A paper submitted to the faculty of
The University of North Carolina at Chapel Hill
in partial fulfillment of the requirements for the degree
Master of Public Administration

Spring 2013

This paper represents work done by a UNC-Chapel Hill Master of Public Administration student. It is not a formal report of the School of Government, nor is it the work of School of Government faculty.

Executive Summary

North Carolina evaluates its public school teachers using both principal observations and SAS Institute’s EVAAS data-driven model of student growth. Literature on other performance evaluation systems shows that correlations between objective and subjective measures of teacher proficiency generally fall between a .17 and a .49. Using evaluation data on 26,260 North Carolina teachers, this study finds a statewide correlation coefficient of .23 between teacher evaluations and EVAAS data—a value consistent with the literature. Nevertheless, certain North Carolina districts exhibit far weaker correlations, and there is evidence that principals give high ratings to the vast majority of teachers and fail to distinguish between different facets of teacher performance. To ensure more accurate and consistent evaluations of teacher performance, this study recommends targeted training and certification of principals on the evaluation process, increased classroom observational time, and a continued effort to incorporate student growth data into teacher evaluations.
Introduction and Research Purpose

Administrators, policymakers, and parents have a vested interest in assessing and improving the performance of public school teachers, but teacher performance is a multi-faceted and complex concept that defies simple evaluation. In order to overcome this challenge, scholars and education professionals have developed several different methods for evaluating teacher performance. Traditional models include “standards-based” systems, which provide administrators with rubrics to use when observing and evaluating teachers. Over the last several years, however, states have been introducing statistical “value-added” models, which use complex mathematical systems to assess teachers’ contributions to student growth in standardized test performance. In fact, according to a 2011 report by the National Council on Teacher Quality (2011:ii), 23 states now incorporate student growth data into teacher evaluations, while 17 have made student growth a “significant” or “preponderant” criterion for evaluating educator effectiveness. Understanding the merits of and relationships between teacher evaluation frameworks will prepare administrators and policymakers to make decisions that will improve student achievement.

The North Carolina Board of Education (2012a) requires principals to evaluate teachers using five Professional Teaching Standards. During a full evaluation (principals may elect to give an abbreviated evaluation in certain circumstances) principals must observe a teacher for at least 45 minutes or one class period, conference with the teacher multiple times, and assign the teacher one of five possible performance ratings for each standard (Board of Education 2011). The Standards and performance levels are shown below.\(^1\)

<table>
<thead>
<tr>
<th>Teaching Standards:</th>
<th>Performance Levels:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Teachers demonstrate leadership</td>
<td>1. Not Demonstrated</td>
</tr>
<tr>
<td>2. Teachers establish a respectful environment…</td>
<td>2. Developing</td>
</tr>
<tr>
<td>3. Teachers know the content they teach</td>
<td>3. Proficient</td>
</tr>
<tr>
<td>4. Teachers facilitate learning for their students</td>
<td>4. Accomplished</td>
</tr>
<tr>
<td>5. Teachers reflect on their practice</td>
<td>5. Distinguished</td>
</tr>
</tbody>
</table>

The Board of Education (2012a) has recently added a sixth standard for teacher evaluations: “Teachers contribute to the academic success of students.” Though Standards Four and Five were already designed to have a “particular focus” (Office of the Governor 2010) on student growth, teachers’ ratings on this sixth standard are determined by three-year rolling averages of actual student growth data observed at the individual and the school level. (For the 2011-2012 school year, Standard Six scores reflected 70% individual and 30% school performance.) On this Standard, teachers may be rated “exceeds expected growth,” “meets expected growth,” or “does not meet expected growth.” In February 2012, the Board (2012b) approved SAS Institute’s Education Value-Added Assessment System (EVAAS), which builds on value-added models described by Sanders, Saxton, and Horn (1997), as the method for calculating Standard Six scores.

Once a principal determines teachers’ ratings on the individual Standards, the teachers then receive an overall assessment of their effectiveness (Board of Education 2012a). Teachers rated at least “accomplished” on all five original Standards and “exceeds expected growth” on Standard Six are considered “highly effective.” Those rated at least “proficient” on Standards One through Five and “meets expected growth” on Standard Six are considered “effective.” Teachers rated “developing” or below on any standard or who do not meet expected growth are considered “in need of improvement” and are placed on growth plans monitored or directed by a school administrator (Board of Education 2011).

The introduction of Standard Six has motivated the Department of Public Instruction to study agreement between EVAAS and the original Professional Teaching Standards. Moreover, the explosion of value-added assessment systems in educator evaluations nationwide has led to significant scholarly interest in both teacher evaluations and teacher value-added data. This study aims to contribute to these discussions.

\(^1\) For a full list of the standards, their constituent elements, and performance levels, see Board of Education (2012a).
**Literature Review**

The only research looking specifically at North Carolina’s Teaching Standards is a validation study completed by Mid-continent Research for Education and Learning (2011). Using a sample of 1,413 teachers in seven pilot districts, the study found a wide range of scores and significant correlation amongst teachers’ performance ratings on the Standards. It also noted that males and probationary teachers had lower evaluation ratings, though it did not control for objective measures of performance.

Beyond the validation study, several areas of academic literature are relevant to North Carolina’s evaluation system: literature on EVAAS, performance evaluations, and the agreement between objective and subjective measures of performance. The following review considers each of these in turn.

Academic discourse on EVAAS and its predecessor, TVAAS, has centered on several questions related to the statistical objectivity of the model. Several authors (Amrein-Beardsley 2008; Kupermintz 2003; Raudenbush 2004; Ladd and Walsh 2002) have claimed that EVAAS and related value-added models insufficiently control for student demographics, biasing the system against teachers with large numbers of low-income or minority students. William Sanders, the creator of the model, and his associates maintain that controlling for past student performance is sufficient to remove any bias stemming from demographic factors (Ballou, Sanders, and Wright 2004; Sanders and Horn 1994). Kupermintz (2003) and Amrein-Beardsley (2008) also point out that teachers with low numbers of students are less likely to stand out as anything other than average in the EVAAS model. Sanders, Saxon, and Horn (1997) admit that this is true, but argue that this practice protects teachers from statistical errors created by small sample sizes. Lastly, Reckase (2004) and Papay (2010) argue that value-added measures are compromised by problems arising from the standardized achievement tests underlying the models, a position consistent with other literature documenting the challenges associated with standardized tests (e.g., Koretz 2002; Marchant, Paulson, and Shunk 2006). Though scholars ultimately disagree about the objectivity and usefulness of EVAAS, it is clear that it is by no means a perfect measure.

Other scholars (e.g., Brophy et al. 1975; Stodolsky 1984; Weisberg et al. 2009; Hill, Charalambous and Kraft 2012) have produced a longstanding body of literature on the difficulties associated with principal observation and evaluation of teacher performance. This literature led some researchers to conclude as early as the 1980s that there is “no support whatever for the widely held belief that the average principal is a good judge of teacher performance” (Medley and Coker 1987). Scriven (1995) points to several sources of inaccuracy in teacher evaluations, including: evaluative criteria that either do not encompass the full range of teacher performance or are irrelevant to teacher performance, small sample size in teacher observations, and altered behavior during the observation. Peterson (2004) adds that principals’ desire to preserve school harmony may undermine their willingness to draw performance distinctions, while style preferences and a lack of subject matter knowledge may taint their perceptions of teacher skill. In terms of teacher observations themselves, a nationwide panel of researchers found that observation ratings vary “considerably from lesson to lesson, as well as from observer to observer” (Measures of Effective Teaching 2012:8) and recommend training and certifying principals on observation instruments and averaging scores across multiple lessons. The same researchers (Measures of Effective Teaching 2013) later note that multiple ratings by multiple observers can greatly increase evaluation reliability, even if some of the additional observations are only 15 minutes. Gerstner and Day (1997) and Lefkowitz (2000) report that a supervisor’s affective regard for the employee can influence performance evaluation ratings. Though surveys of performance evaluation literature (Landy, Shakster, and Kohler 1994; Roth, Bobko, and Huffcutt 2003) find no evidence of systematic bias against a particular race or gender, other studies have examined the interactions of rater and ratee characteristics and have documented significant gender and race effects (Varma and Stroh 2001; Payne 2011; Chung 2001; Elvira and Town 2002).

In light of the problems associated with both objective and subjective performance measures, scholars have begun looking at correlations between the two. In meta-analyses of literature on performance evaluation in other industries, Heneman (1986) and Bommer et al. (1995) found moderate levels of
correlation between objective and subjective measures, with average coefficients of .27 and .389, respectively. Similar studies of teacher performance (Medley and Coker 1987; Milanowski 2004; Jacob and Lefgren 2008; Holtzapple 2003) placed correlation coefficients for teacher evaluations and value-added scores somewhere between .17 and .49 (weak to moderate), depending on data and methodology. Other studies (Rockoff and Speroni 2010; Gallagher 2004; Schacter and Thum 2004) found evidence of agreement between value-added measures and performance reviews, but did not report a coefficient. Finally, in a study of performance data from a Nevada school district, Kimball et al. (2004) found only mixed evidence of correlation between the subjective and objective measures. Though establishing an exact target for such correlations may be a matter of personal preference or arbitrary decision making—Medley and Coker (1987:246) call a coefficient of .26 “too small to be useful,” for example—it is clear that student growth data and teacher evaluations should correlate to at least some extent.

**Methodology**

The Department of Public Instruction provided 2011-2012 evaluation data for all North Carolina teachers, including ratings on Standard One and Standard Four for career status teachers receiving abbreviated assessments and Standards One through Five for all others. The Department also provided value-added scores for all teachers assessed using EVAAS. The final resulting data set reflected principal evaluations and EVAAS data on 26,260 teachers.

Composite evaluation and student growth scores were needed to facilitate a comprehensive analysis. Each teacher was assigned an average performance evaluation score reflecting the mean of their available performance ratings. Since EVAAS student growth scores are not reported on a consistent scale, values for each unique class or class/grade combination (e.g., Algebra I or Grade 5 English) were standardized such that they had a mean of zero and a standard deviation of one. All EVAAS scores available for each teacher were then averaged to create a composite score.

A statewide correlation coefficient was computed for the composite teacher evaluation and EVAAS data sets. However, because Milanowski (2004), Jacob and Lefgren (2008), and Holtzapple (2003) used district-level datasets for their calculations, correlations were also calculated at the district level. In addition to maintaining consistency with existing literature, examining district correlations avoided errors that might arise from inconsistent application of the Professional Teaching Standards. Finally, in order to test the “particular focus” (Office of the Governor 2010) of Standard Four on student growth, correlations between Standard Four ratings and EVAAS data were calculated at the state and district levels.

**Statistical Results**

The statewide correlation between student growth measures and composite performance evaluation ratings was a .23. This value is consistent with those observed in the teacher performance literature, though it lies closer to the bottom of the range (.17) than the top (.49), as shown in the chart below. The correlation coefficient for EVAAS scores and Standard Four ratings was just slightly lower at a .22.

At the district level, coefficients for composite ratings were slightly higher, with an average value of .26. Individual districts varied in the strength of their correlations from -.14 to .61. Extreme values tended to be concentrated in districts with low numbers of teachers, likely as a result of statistical error resulting from a small N. In the 65 districts with at least 100 teachers in the data set, correlations ranged from .08 to .52. As with the statewide data, the correlation between EVAAS scores and Standard Four ratings was a .22.

---

2 See Appendix 1 for a full list of correlations for school districts with more than 100 teachers in the data set.
comparatively weaker .23. Values ranged from a -.17 to a .62 for all districts and from .06 to .44 for districts with an N of at least 100.

Discussion

These results show that the North Carolina Professional Teaching Standards are comparable to other rubrics used for evaluating teacher and employee performance. Correlation coefficients computed in this study lie within the range of coefficients found in the teacher performance literature and are slightly below average correlations of subjective and objective measures of employee performance observed in meta-analyses conducted by Heneman (1986) and Bommer et al. (1995). This study therefore confirms a certain degree of agreement between North Carolina teachers’ performance evaluations and student growth, though that agreement may be slightly weaker than that observed with other instruments.

Nevertheless, given scholarly and professional dissatisfaction with teacher evaluation frameworks in general, falling within the range of correlations observed nationally is not necessarily evidence of ultimate success. In fact, several factors may be undermining the agreement between the Teaching Standards and student growth. First, it is possible that the Standards incorporate facets of teacher performance not measured by student growth data. Unfortunately, however, weaker levels of correlation between EVAAS data and Standard Four ratings (discussed more fully later) show that the comprehensiveness of the Standards is not a sufficient explanation. Second, the wide variations in district correlations may be an indication that many principals may not have the training necessary to take full advantage of the Standards. Third, evaluation scores are high and narrowly distributed, as shown in the chart below. More than 50% of North Carolina teachers have an average evaluation score of 4 or higher. Moreover, the 100 teachers with the lowest student growth scores averaged a 3.4 on their performance evaluations while the 100 teachers with the highest student growth averaged only a 4.1. It is therefore likely that principals face incentives to adjust evaluations to preserve morale. Fourth, as noted in scholarly literature, a single 45-minute observation may be insufficient to adequately understand teacher performance. Finally, principals may bias evaluations based on race, gender, experience, or training.

The statistical results also show that, despite statements that student growth is a “particular focus” of Standard Four, such a focus is not reflected in teachers’ Standard Four ratings. Correlations between EVAAS data and Standard Four were actually weaker than those between EVAAS scores and composite evaluation scores at both the state and district levels. There are two possible explanations for this. First, principals may not treat Standard Four as a measure of student growth due to insufficient training on the Standards or due to a desire to give more or less consistent scores across the Standards. Second, principals may recognize the individual focus intended for each standard, but may view teacher performance holistically rather than as a combination of several constituent parts. In either case, there is a clear disconnect between the State’s intentions for Standard Four and the way that it is applied “in the field.”

Finally, this study shows that there is significant variation between school districts in the agreement between student growth and teacher performance evaluations. Though random error in the data undoubtedly accounts for some of the differences, levels of correlation fluctuated greatly even amongst districts with a large number of teachers in the data set. Thus, even accounting for random error, there are significant variations from district to district in the agreement between student growth and teacher performance evaluations. Such variations likely reflect differences in principal training, district policies, and other outside factors.
Policy Recommendations

Based on these findings, the Board of Education and the Department of Public Instruction should take four significant steps towards strengthening North Carolina’s teacher evaluation framework.

1. **Implement focused training and certification.** Many districts exhibit little or no correlation between teacher evaluations and student growth and there is evidence that principals may misunderstand the intention of Standard Four. The State should use training to address these issues and consider assessing principals’ application of the Professional Teaching Standards.

2. **Encourage a broad distribution of scores.** Study data show that teacher evaluation scores are high and narrowly distributed. Principals must give a broad range of scores so they can identify those teachers who are excelling and those who may need additional help.

3. **Modify observation requirements.** Research shows that a single observation is insufficient to establish a valid and reliable understanding of teacher performance. North Carolina should incorporate ratings from multiple observers in several different observation sessions (some of which could be as short as 15 minutes).

4. **Keep Standard Six and keep it distinct.** Neither EVAAS nor the Professional Teaching Standards are a perfect method for evaluating educator performance. Using both measures in tandem will allow North Carolina to achieve more accurate results with its teacher evaluation system. Additionally, though evaluation scores should be correlated with student growth, they should also incorporate other significant facets of teacher performance not captured by EVAAS.

Limitations and Areas for Further Study

Two primary factors limit the results of this study. First, many teachers in the data set were evaluated only on Standards One and Four, making their average scores qualitatively different than the mean of all five Standards used for other teachers. However, a different composite evaluation score constructed only from Standards One and Four showed similar correlations to EVAAS data, implying that differences in the number of available ratings did not affect study results. Second, student growth figures provided by EVAAS are not an exact measure of teacher effectiveness. Because statistical error in value-added models is drastically reduced when multiple years of data are included (Schochet and Chiang 2012), North Carolina evaluates teachers only using a three-year rolling average of EVAAS data (Board of Education 2012a). Due to recent changes in the State’s EVAAS implementation and because other literature did not use three-year averages this study used data only from the 2011-2012 school year. The statistical error inherent in this data, combined with other drawbacks to EVAAS discussed earlier, reduce the precision of the student growth estimates.

This study also highlights two promising areas for future study. First, researchers should seek to better understand the reasons that teacher evaluations and student growth data differ in those cases where they are very weakly correlated. Similarly, the State should seek to understand the particular issues that apply in North Carolina. Researchers should consider factors including inadequate training, insufficient time observing in the classroom, and evaluator bias. Second, researchers should investigate the “Standard Four” problem to determine whether principals misunderstand the evaluation instrument, or whether principals view teacher performance holistically and have trouble identifying its constituent parts.

Conclusion

This study finds that correlations between EVAAS and the Professional Teaching Standards are consistent with correlations of other subjective and objective measures of teacher performance. Nevertheless, certain school districts are exhibiting weak correlations between the two measures, while results statewide show that principals are not necessarily differentiating between the individual Standards to an appropriate degree. Investigating and overcoming these challenges will make the Teaching Standards a stronger tool for evaluating and reporting the performance of North Carolina teachers.
Appendix 1 – District Level Correlations

| District Correlation (Districts Where N>100) |
|-----------------|-----------------|-----------------|
| Alamance-Burlington | 414 | 0.204 | Edgecombe       | 130 | 0.495 | Pasquotank      | 115 | 0.392 |
| Alexander        | 112 | 0.124 | Forsyth         | 931 | 0.318 | Pender          | 157 | 0.298 |
| Beaufort         | 163 | 0.302 | Franklin        | 147 | 0.322 | Pitt            | 472 | 0.328 |
| Bladen           | 101 | 0.289 | Gaston          | 558 | 0.270 | Randolph        | 347 | 0.344 |
| Brunswick        | 250 | 0.162 | Granville       | 155 | 0.525 | Richmond        | 145 | 0.216 |
| Buncombe         | 513 | 0.249 | Guilford        | 1300| 0.260 | Robeson         | 348 | 0.208 |
| Burke            | 253 | 0.317 | Harnett         | 326 | 0.449 | Rockingham      | 259 | 0.183 |
| Cabarrus         | 520 | 0.161 | Haywood         | 152 | 0.134 | Rowan-Salisbury | 380 | 0.295 |
| Caldwell         | 253 | 0.104 | Henderson       | 259 | 0.364 | Rutherford      | 182 | 0.175 |
| Carteret         | 170 | 0.226 | Hoke            | 137 | 0.194 | Sampson         | 176 | 0.156 |
| Catawba          | 321 | 0.212 | Iredell-Stateville | 410 | 0.229 | Scotland        | 120 | 0.358 |
| Chapel Hill-Carrboro | 162 | 0.082 | Johnston        | 620 | 0.232 | Stanly          | 184 | 0.168 |
| Chatham          | 158 | 0.268 | Lee             | 196 | 0.203 | Stokes          | 146 | 0.195 |
| Cleveland        | 307 | 0.264 | Lenoir          | 188 | 0.273 | Surry           | 165 | 0.173 |
| Columbus         | 131 | 0.296 | Lincoln         | 244 | 0.261 | Union           | 738 | 0.233 |
| Craven           | 291 | 0.189 | McDowell        | 136 | 0.191 | Vance           | 127 | 0.398 |
| Cumberland       | 953 | 0.226 | Mecklenburg     | 2251| 0.274 | Wake            | 2396| 0.132 |
| Dare             | 101 | 0.255 | Moore           | 243 | 0.210 | Wayne           | 292 | 0.293 |
| Davidson         | 375 | 0.242 | Nash-Rocky Mount | 327 | 0.303 | Wilkes          | 205 | 0.282 |
| Davie            | 130 | 0.108 | New Hanover     | 409 | 0.200 | Wilson          | 224 | 0.382 |
| Duplin           | 182 | 0.258 | Onslow          | 430 | 0.156 | Yadkin          | 129 | 0.127 |
| Durham           | 444 | 0.255 | Orange          | 137 | 0.288 | Statewide       | 26260| 0.227 |

![District Correlations (All Districts, Rounded)](image)
Bibliography


