

Technical Report for the Effectiveness Study, 2008 - 2009 Commissioned for the Tennessee Teacher Quality Reforms

SECTION 1: INTRODUCTION

The Tennessee Teacher Quality Reforms initiative aims to improve student achievement and educational attainment in the state as a part of the state mandate to “develop a report card or assessment on the effectiveness of teacher training programs” (TCA 49-5-108). A key part of this goal will be realized via state and local programs focused on new teachers in terms of the recruitment, selection, preparation and support for these new teachers. The State of Tennessee asked SAS® EVAAS® to compare the teaching effectiveness of recent licensure recipients from various teacher preparation institutes to the effectiveness of other teachers in the state.

Thus, the goals of the effectiveness study were:

- To identify any university that tends to produce beginning teachers who are highly effective as well as to identify any university that tends to produce beginning teachers who are very ineffective
- To determine if a university is above or below the reference distribution with a fair and reliable statistical test

The importance of identifying such teacher training programs is evident in comparing the mean teacher NCE gain between highly effective teachers and highly ineffective teachers. This measure represents the average gain in learning for students. The chart below shows the mean teacher NCE gain for both the highest and lowest quintiles of teachers in the state for various subjects.¹ The difference between the two groups reveals the substantial impact on student progress in terms of a student having a teacher from the highest or lowest quintile.

¹ How the quintiles were selected is described later in this report.

Chart 1: Mean Teacher NCE Gains²

TCAP Subjects	Quintiles	
	Low	High
Math	-5.228	4.734
Reading/Language	-2.478	3.198
Science	-4.560	4.684
Social Studies	-4.820	4.854

In realizing the goals to assess teacher training programs, the effectiveness study also sought to provide a fair, rational method of comparison that is statistically sound, easy to interpret, and useful to both policymakers and the public. This was accomplished by examining the difference between the beginning teachers from each institution and two reference groups described in Section 4. This report is a technical document that explains these analyses in detail. This report does not include any results to the effectiveness study.

SECTION 2: KEY ELEMENTS OF THE TWO ANALYSES

The two analyses chosen to address the effectiveness study's goals used the same underlying data. This section describes what data were used, why and how they were used in the analyses, and the applied definition of effectiveness.

Data Used in the Effectiveness Study

The only teachers included in these two analyses were those who have value-added data from the Tennessee Value-Added Assessment System (TVAAS), which is "a statistical system for educational outcome assessment which uses measures of student learning to enable the estimation of teacher, school, and school district statistical distributions" (TCA 49-1-603). TVAAS has been a part of state statute since 1992, and its use results in an extensive and useful statewide database on educational attainment of Tennessee students.³ The longitudinal, multivariate, mixed-model methodology of TVAAS produces more reliable estimates with less bias than other more simplistic models, an opinion recently corroborated by researchers at RAND.⁴ TVAAS has produced

² Appendix 1 contains two additional charts similar to Chart 1, and they show the mean teacher NCE gain for new teachers.

³ More specific information on TVAAS methodology is available online at <http://www.sas.com/govedu/edu/sanderssaxtonhorn.pdf>

⁴ McCaffrey, D. F., Han, B. and Lockwood, J. R. (2008). *From Data to Bonuses: A Case Study of the Issues Related to Awarding Teachers Pay on the Basis of the Students' Progress*. Paper presented at the conference on Performance

teacher effect estimates since 1996, and these estimate a teacher's impact on student learning, as measured by students' performance on standardized tests, such as TCAP, Gateway and End-of-Course.

The teacher effect estimates were based on the TCAP subject tests in math, reading/language arts, science and social studies in grades four through eight as well as the high school End-of-Course and Gateway tests. Thus, teachers who teach non-tested subjects were not included in the analyses.

An additional group of teachers who were not included in the study were those who teach primarily special education students or students with low attendance records. This is because state statute prohibits the use of these students in value-added analysis (TCA 49-1-606).

SAS received two files from the State of Tennessee linking all teachers who had received their licensure from one of 39 Tennessee teacher preparation institutions to their respective institution of licensure.⁵ One file contained teachers who were licensed through the traditional route during the years 2002 – 2008 while the other file contained teachers who were licensed through the alternative route during the years 2003 - 2009. The timeframe was selected due to the study's focus: the effectiveness of teacher training programs in preparing beginning teachers, with the implicit assumption that other factors beyond the licensing institution could become quite influential in later years. At the request of the State of Tennessee, the definition of "beginning" teacher is those with 1 – 3 years of experience.

How the Data Were Used

Because individual teacher effects are private by state statute (TCA 49-1-606), the effectiveness study reported teacher effect data by group (subject, institution, type of licensure, etc.) so that the privacy of the teachers was not compromised. The grouping also increased the counts for each particular group so that fair comparisons could be made among teacher training programs since most institutions do not produce many teachers in a given subject/grade each year. More specifically, the study considered all grades in each subject together. In order for an institution

Incentives: Their Growing Impact on American K-12 Education, February 28-29, National Center on Performance Incentives at Vanderbilt University's Peabody College: "Multivariate mixed model methods and fixed effects methods with shrinkage tend to provide estimates that appear to have relatively less noise and relatively less bias. Performance

measures from both methods tend to have strong cross-year correlation within teacher, weak correlation with students'

prior achievement, and relatively few teachers with small classes ranked in the extremes of the sample" (p. 37).

⁵ See Appendix 2 for a list of the teacher training programs.

to be included in the analysis for a particular subject, a minimum of five teachers from that institution were required. Results were reported for each type of licensure as well as for both types together.

Due to the emphasis on beginning teachers and the preparation received by their institutions, the effectiveness study utilized one-year estimates of teacher effectiveness from the year 2008 - 2009. More specifically, the *t-value* of the teacher effect was used as the basis of comparison rather than the teacher effect itself or the teacher gain.⁶ This solved three major problems, two of which apply specifically to TCAP tests.

First, using a measure based on the teacher effect rather than the teacher gain overcame issues relating to random assignment. Teachers from different institutions are not randomly assigned to their school districts; geography typically plays a role in the assignment. Because the TCAP tests utilize a value-added teacher effect that is centered on the district gain, an institution with a disproportionate number of their teachers in a district with either a very high or low gain could have a skewed comparison if teacher gain was used as the measure for evaluating teachers. By using a measure related to the teacher effect, the impact of the disproportional location of teachers from different teacher training programs was removed. Note, the district centering was not an issue for Gateway and End-of-Course tests because they utilize a value-added teacher effect centered on the average teacher in the *state* of Tennessee.

As a second advantage, using the *t-value* of the teacher effect, instead of the teacher effect alone, enables equitable comparisons across multiple grades, which was necessary for the reasons stated above. Because teacher effects are shrinkage estimates (BLUPs) in TVAAS methodology, they shrink back towards zero. In practice, this means they shrink back towards the district gain since the teacher effects are centered on the district gain. Because teacher variance components vary among grades, there are different amounts of shrinkage among different grades. For example, higher grades typically have less shrinkage. Thus, if one institution produces more teachers in higher grades than other institutions, then that institution could have an unfair advantage in any comparison because its teacher effects would likely have less shrinkage. However, as the shrinkage of any teacher effect increases, the standard error of the teacher effect decreases. Therefore, using the *t-value* of a teacher effect allowed a more fair comparison among teachers in different grades than using the teacher effect itself. Again, this issue did not

⁶ Teacher effect measures teacher effectiveness relative to the district average gain and is part of the solution to the mixed model equations for TCAP subjects. The *t-value* of the teacher effect is defined as the teacher effect divided by its standard error in all subjects. Teacher gain is defined as the teacher effect added to the district gain.

apply to Gateway and End-of-Course tests. However, for consistency as well as for the reason outlined below, the t-value of teacher effect is used for the high school subjects as well.

Finally, the use of the t-value of the teacher effect created a fair measure because teachers with very little data tend to have larger standard errors that shrink their measure towards zero. As a result, the use of the t-value promoted the use of teachers with sufficient data for evaluation. This benefit applies to TCAP tests as well as the Gateway and End-of-Course tests.

Definition of Effectiveness in the Study

At the request of the State of Tennessee, highly effective teachers were defined as those teachers in the highest quintile of the state distribution for their subject and grade, as measured by the t-value of the teacher effect. Likewise, highly ineffective teachers were defined as those teachers in the lowest quintile of the state distribution of teacher effect t-values for their subject and grade. The subject/grade combination was used as the basis of analysis so that teachers within any given subject/grade would not have any unfair advantage over any other subject/grade group. As demonstrated in the chart on page one, the study's emphasis on the highest and lowest quintiles is important because the difference in teacher gains between these two groups is substantial.

SECTION 3: IDENTIFYING INSTITUTIONS THAT TEND TO PRODUCE EITHER HIGHLY EFFECTIVE OR VERY INEFFECTIVE TEACHERS

The key elements discussed in Section 2 were then used to address the first goal of the study: identify whether an institution tends to produce more or less of these extreme teachers. To do so, the effectiveness study assessed the percentage of teachers from each institution in either the highest or lowest quintile, as measured by the t-value of their teacher effects. These percentages were compared to the state distribution and tested for statistical significance. In this way, policymakers can assess the effectiveness of teacher training programs in the state.

Defining the Quintiles and Percentages

As described in the previous section, quintiles used for this analysis were based upon the statewide distribution of the t-value of teacher effects from 2008 - 2009 value-added data. By definition, if an institution produced the same percentage of teachers as the state in each of these quintiles, then that institution would have 20% of its teachers in the quintile.

For each institution, the number of teachers in each of these quintiles was compared to the institution's total number of teachers, thus showing the percentage of teachers from a particular teacher training program in either the highest or lowest quintile.

Defining the Model

The difference between the institution's percentage of teachers in the extreme quintiles and the state's percentage was then tested for statistical significance in order to verify that the institution did tend to produce either highly effective or very ineffective teachers relative to the state population. Upper and lower quintiles were analyzed separately to avoid the inclusion of the middle quintile teachers (quintiles 2 – 4) since this latter group was not the focus of the effectiveness study. If an institution had less than five teachers in a subject/grade group, then they were not included in this analysis.

The model for this analysis utilized the binomial distribution to assess statistical significance, with a null hypothesis that the institution distribution is the same as the state distribution. More specifically, in the upper quintile analysis, a teacher was identified as either in the upper quintile or not. The number of teachers who fall into the upper quintile is distributed as a binomial distribution with success probability of 0.20 and the number of trials as the total number of teachers from that institution. Each institution had a certain percentage of teachers who fell into the upper quintile. The exact probability of this can be computed, assuming the null hypothesis, to provide a statistical test for whether or not the true probability of success is different from 0.20. A level of 0.10 was used to determine significance. Thus, if the probability was less than 0.10 of observing a value equal to or more extreme than the percentage of teachers in this quintile for a given institution, then the null hypothesis was rejected: there is sufficient evidence to show that the institution had a probability of producing teachers in the upper quintile that was either more or less than 0.20. The description of this analysis applied to the lower quintile analysis as well.

The tests described above provide a statistical comparison between each institution and the state distribution with respect to the percentage of teachers being produced that are highly effective or very ineffective.

Interpreting the Analysis

While the lower quintile analysis was the same as that for the upper quintile, the interpretation of the test for each quintile is different. For the lower quintile, it is better to have less than 20% of an institution's teachers in that quintile. Conversely, for the upper quintile, it is better to have more than 20% of an institution's teachers in that quintile.

If an institution has a statistically larger percentage of upper quintile teachers than the state distribution, then it tends to produce more highly effective teachers. Likewise if an institution has a statistically smaller percentage of lower quintile teachers than the state distribution, then it tends to produce less ineffective teachers. Teacher training programs with these qualities are doing a good job at producing beginning teachers. The reverse will also show teacher training programs that are doing a poor job at producing beginning teachers.

SECTION 4: DETERMINING IF A UNIVERSITY IS PRODUCING BEGINNING TEACHERS EITHER ABOVE OR BELOW THE REFERENCE DISTRIBUTION

The percentage of teachers from each institution who were either in the highest or lowest quintile provides very useful information to the effectiveness study, but a direct comparison of the teachers from one institution to a reference population would add to an understanding of how a teacher training program is performing overall. The mean t-value of the teachers has a direct relation to value-added analysis, which can enhance understanding among Tennessee's policymakers, educators, and public. Thus, the key elements discussed in Section 2 were then used to address the second goal of the study: determine if a university is above or below the reference distribution with a fair and reliable statistical test. This section describes how such an application was utilized.

Defining a Reference Population

The effectiveness study compared the performance of beginning teachers from the 39 institutions to the performance of teachers in a reference population. In this part of the study, there were two reference populations used for comparison, and they are each described below.

In the first set of analyses, the reference population was a control group that included any teacher who had more than three years of experience from the statewide distribution of teacher value-added data in the 2008 – 2009 school year. Using this reference population, the beginning teachers were compared by institution to these veteran teachers. In this set of analyses, the reference population included all types of licensure.

In the second set of analyses, the reference population was a control group that included beginning teachers linked to the 39 Tennessee institutions. If an institution did not have at least five teachers in a particular subject, then all teachers from that institution were removed from that subject's analyses. In this set of analyses, the reference population and comparison group had the same type of licensure, i.e.,

traditionally licensed beginning teachers were compared to other traditionally licensed beginning teachers.

Defining the Model

The calculation of the mean t-values of the teacher effects utilized a one-way ANOVA model with institution as the fixed effect. In addition to the 39 institutions of higher education used in the model, the institution effect comprised two other levels: (1) teachers with more than three years of experience and (2) any teacher who had three years or less of value-added data with an *unknown* institution of certification. This last group of teachers could include, for example, any teachers who came from other states or who may have been teaching non-tested subjects. For these reasons, they were included as a separate level of the effect. The three types of the institution effect provided the analyses with three distinct and possibly quite different groups of teachers. As such, the model allowed for different levels of variation in each group to ensure that an appropriate statistical test was utilized for each reference population.

As a first comparison, each teacher training program was compared to the veteran teachers in the model, provided that an institution had five or more teachers in that particular subject. The difference of the estimated mean teacher t-value of effects for each comparison was tested for significance.

As a second comparison, each teacher training program was compared to the beginning teachers. More specifically, each institution mean was compared to the mean of all of the institution means, with each institution weighted the same. The number of teachers for every institution was not a part of this weight since it would cause a small number of institutions to dominate the mean. This method of weighting ensured a more fair comparison among institutions. Again, if an institution had fewer than five teachers, then its data were removed from the analysis due to an insufficient number of teachers for a reliable statistical estimate.

As a third comparison, the difference between the two reference populations was considered to determine if the beginning teachers from the institutions were significantly different from the veteran teachers in Tennessee. More specifically, the mean of veteran teachers was compared to the mean of institution means for beginning teachers, provided that the beginning teachers' institution had at least five teachers in the subject being analyzed.

Index for Comparison

For ease of interpretation and utility for comparing the teacher training program, an index was created, based on the mean t-value of teacher effects. In the calculation of this index, each institution mean was compared with the mean of the reference population.

Each difference was between an individual teacher training program and the reference group, which represented either the veteran teachers or the beginning teacher subset.

The index analyses sought to present a balanced assessment of the net effectiveness of each teacher training program by showing how average teachers from each program would compare to the reference population. If any difference between the institution and reference mean is positive, then the institution mean is greater than the reference population mean t-value of teacher effects. A significant positive number indicates that a teacher training program has produced beginning teachers with statistically significantly larger mean t-values as compared to the reference population in terms of a teacher's mean t-value of effects in 2008 - 2009. A level of 0.10 was used to test statistical significance. These comparisons were made by type of licensure as well as by both types together for institutions that had sufficient data.

Interpreting the Indices

In the TCAP subjects, the mean t-value of teacher effects for each group (i.e., subject/grade combination for a particular institution) is a meaningful comparison that does not confound the district distribution of teachers and is also interpretable in NCE value-added teacher gains. The mean t-value can be interpreted as follows: on average, teachers in this group have estimated teacher gains that are X number of standard errors away from their district's mean NCE gain, where X represents the index for comparison. In other words, teachers in that group have sufficient data to show their estimated teacher gain is either above or below their district's mean NCE gain by the reported factor.

In the high school subjects, the mean t-value of teacher's effects is also a meaningful comparison across the state of Tennessee. The mean t-value can be interpreted as follows: on average, teachers in this group have estimated teacher effects that are X number of standard errors away from the average teacher effect in the state of Tennessee, where X represents the index for comparison. In other words, teachers in that group have sufficient data to show their estimated teacher effect is above or below the average teacher effect in the state of Tennessee.

Thus in both cases, an institution producing beginning teachers with significantly better t-values of teacher effects will have a positive impact

on student progress. Ideally, new methods of training at the institutions enable beginning teachers to outperform existing teachers.

SECTION 5: REPORTING THE RESULTS OF THE EFFECTIVENESS STUDY

The effectiveness study results present the number, percentages, and index measures associated with each of the 39 Tennessee institutions by subject as long as that teacher training program has sufficient data. If the percentage or index measure is statistically significant from the statewide average at the 90% confidence level, this will be noted. Results were presented by institution for each type of license as well as for both types together.

Appendix 1: Mean Teacher NCE Gain for Beginning Teachers

Chart 2: Mean Teacher NCE Gains for Beginning Teachers with 1-3 Years Experience

TCAP Subjects	Quintiles	
	Low	High
Math	-5.500	4.575
Reading/Language	-2.711	2.879
Science	-4.953	4.252
Social Studies	-5.057	4.326

Appendix 2: List of Participating Institutions

Aquinas College
Austin Peay State University
Belmont University
Bethel College
Bryan College
Carson-Newman College
Christian Brothers University
Crichton College
Cumberland University
David Lipscomb University
East Tennessee State University
Fisk University
Free-Will Baptist Bible College
Freed-Hardeman College
Johnson Bible College
King College
Lambuth University
Lane College
LeMoyne Owen College
Lee College
Lincoln Memorial University
Martin Methodist College
Maryville College
Middle Tennessee State University
Milligan College
Rhodes College
Southern Adventist University
Tennessee State University
Tennessee Technological University
Tennessee Wesleyan College
Trevecca Nazarene University
Tusculum College
Union University
University of Memphis
University of South
University of Tennessee, Chattanooga
University of Tennessee, Knoxville
University of Tennessee, Martin
Vanderbilt University