Value Added Assessment of Teacher Preparation in Louisiana: 2004-2005 to 2006-2007

George H. Noell, Ph.D.
Bethany A. Porter, M.A.
R. Maria Patt, M.A.
Amanda Dahir, M.A.

Department of Psychology
Louisiana State University

November 17, 2008

Acknowledgements

This report is based upon data provided by the Division of Planning, Analysis, and Information Resources of the Louisiana Department of Education. The authors would like to thank Allen Schulenberg, Kelvin LaCroix, Sam Pernici, and David Elder (Division Director), without whom this work would not have been possible. Additional data were provided by the Louisiana Board of Regents.

Any errors, omissions, or misstatements contained herein are entirely the responsibility of the authors. Any conclusions proffered are the responsibility of the authors and do not reflect the views of the Louisiana Department of Education or the professionals from that organization who provided professional guidance and technical assistance.

This work was supported by award CT-06/07-VAA-01 from the Louisiana Board of Regents.

Table of Contents

Abstract

Value Added Assessment of Teacher Preparation

Analyses were conducted examining the degree to which recent graduates of specific teacher preparation programs were associated with increased or decreased educational attainment of students taught by their graduates as compared to experienced teachers. Work began with the construction of a large multivariate longitudinal database linking many data points. This was followed by a model development phase in which hierarchical linear models were developed to predict student achievement based upon prior achievement, student demographic factors, and classroom level covariates. The models nested students within teachers and teachers within schools. Separate models were developed for each content area. These models were used to assess the efficacy of teacher preparation programs. Analyses were conducted across a pooled data set spanning the academic years 2004-2005 to 2006-2007. Due to the timing of teacher preparation program (TPP) redesign and the meaning of the data relevant to current programs, results are limited to redesigned alternative teacher preparation programs. Data regarding traditional undergraduate programs will become available as sufficient graduates of their new programs enter the teacher workforce. Effect estimates were identified at all five performance bands that were developed to describe teacher preparation programs. Some consistency in TPP effects was evident with three programs clustering in the top of the distribution and two clustering in the bottom across content areas. Results were generally consistent for TPPs at the level of performance bands across the 2007 and 2008 reports, with one TPP exhibiting a decline of one level across all content areas. ACT scores were sufficiently consistent across programs that selection as indexed by ACT scores could have little explanatory value for program differences. The ACT score in mathematics taken prior to teacher preparation was a statistically significant predictor teaching effectiveness in mathematics. Teachers who were not certified in the content they were teaching were less effective than those who were content certified.

Technical Report:
Value Added Assessment of Teacher Preparation in Louisiana:
2004-2005 to 2006-2007

I. Introduction

This report describes the results of the *Value Added Assessment of Teacher Preparation Project* (VAA-TPP) for the academic years 2004-2005 to 2006-2007. These analyses build upon results reported previously in Noell (2006) and Noell, Porter, and Patt (2007). This study extends those studies by adding data from the 2006-2007 school year. The VAA-TPP project is a research study housed in the Department of Psychology at Louisiana State University. The VAA-TPP is building longitudinal databases linking students across years and linking those students to their teachers in core content areas. The project has been examining the feasibility and initial results of using this data system to examine the impact of teacher preparation programs (TPP).

The goal of the research is to develop a model for the assessment of TPP as pathways into the teaching profession. At this stage in its development the VAA-TPP examines the average impact of new teachers from specific preparation programs. The research team does not have sufficient data to examine the differential effects of TPP in domains such as recruitment, admissions, content preparation, pedagogical preparation, field experiences, screening for graduation, or transition into the workforce. A separate statewide research team led by Dr. Jeanne Burns that includes representatives from all TPPs in Louisiana is currently collecting data examining the process of teacher preparation. Over the next year these data will be integrated into the longitudinal data that are the basis of this report, and will provide the foundation for initial efforts to examine the process of teacher preparation within the VAA-TPP.

In the context of this report, value added analysis describes the use of demographic and prior achievement data to estimate expected outcomes for students in a specific content domain (e.g., mathematics) based on a longitudinal data set derived from all students who took state mandated tests in grades 3 to 9. The assessment uses a relatively complex model that includes the grouping of students within classrooms and classrooms within schools. The model then examines the degree to which students who are taught by new teachers from specific TPPs compare to other students after controlling for prior achievement and demographic factors. This information is used to estimate the degree to which new teachers' effectiveness is differentially associated with having entered teaching through a TPP.

The estimation of educational effects within complex longitudinal models that can accommodate the correlation of errors that emerge due to the nesting of students within classrooms and schools is a complex and emerging literature base that is beyond the scope of this technical report (see for example Ballou, Sanders, & Wright, 2004; Goldhaber & Brewer, 1997; Hill, Rowan, & Lowenberg, 2005; Hong & Raudenbush, 2008; McCaffrey et al., 2003; Todd & Wolpin, 2003; Wayne & Youngs, 2003). This technical report will summarize the findings of the analyses through 2006-2007. Additional work is underway that examines how findings in Louisiana fit into the broader literature relevant to the assessment of TPP and instruction.

Prior Work

The current study builds on initial pilot work from a sample of opportunity of 10 school districts within Louisiana (Noell, 2004; Noell, 2005). Initial pilot investigations suggested that it may be possible to detect differences between teacher preparation programs and that those differences might be relatively stable across years.

Subsequent analyses were conducted in 2006 and 2007 based on a statewide longitudinal database (Noell, 2006; Noell, Porter, & Patt, 2007). These studies developed a general assessment approach within the framework of hierarchical linear models (HLM; McCulloch & Searle, 2001; Raudenbush & Bryk, 2002) that nested students within teachers and teachers within schools. The studies found some differentiation among preparation programs along with considerable overlap of TPP effect estimates. HLMs capture the natural nesting of students and classrooms within schools and as a result permit correlation of error terms within nested units. This permits modeling of variables at the student, classroom, and school level in a methodologically appropriate manner. The nesting structure also permits a model in which effects can be appropriately linked through the hierarchy such as the effect of schools upon teachers who in turn affect students.

The prior work examined a number of specific issues in the specification of the assessment models. For example, based on examination of estimated teacher effects by years of experience cohorts, new teachers were defined as first *and* second year teachers (Noell et al., 2007). Additionally, the minimum standard for reporting results for an individual university was set at 25 observations of teacher/year outcomes based on an examination of the ratio of variance within program estimates to the variance between programs relative to the number of graduates (see Noell et al. for a detailed discussion).

One of the most important modeling conventions adopted within the prior work was the decision to use a single year covariate adjustment approach for modeling student achievement (Noell, 2006; Noell et al., 2007). This approach uses five achievement test scores from the prior year combined with more than 12 demographic variables to predict current year achievement. While these models have extensive specifications that account for a substantial portion of the variance in student achievement, they do not capitalize on the analytic power and elegance of multiyear achievement trajectories for students across multiple teachers (see Nye, Konstantopoulos, & Hedges, 2004; McCaffrey et al., 2003; McCaffrey et al., 2004; Sanders & Horn, 1998; Todd & Wolpin, 2003).

The decision to use a covariate adjustment approach was guided by two considerations. First, the covariate adjustment models were able to account for a substantial portion of the variance in achievement, suggesting they may be sufficient for this type of assessment. Second, multiyear repeated observation models assume that the quantity that is being observed across years is an unchanging one-dimensional scale such as dollars or truly vertically aligned educational tests (Matrineau, 2006; Seltzer et al., 1994). Although there can be considerable debate about the degree to which vertical scaling is actually achieved or is achievable in educational assessment over wide grade spans (see Matrineau et al., 2007; Reckase, 2004), a plausible argument cannot be made that Louisiana's assessments are vertically aligned. The tests are aligned to the content standards for each grade and as a result are an assessment of the blueprint of instruction. However, that means that the specific content and weighting of the content shifts considerably from one year to the next. This is particularly striking in science and social studies where some years are thematically focused (e.g., life science or Louisiana history). Interested readers can examine http://www.doe.state.la.us/lde/saa/2273.html for a description of Louisiana's assessment content by grade level. A covariate adjustment model can be built upon relatively modest assumptions regarding the measurement properties of the tests that contribute to them (see Matrineau et al., 2007; Reckase, 2004; Seltzer, Frank, & Bryk, 1994) and these assumptions appear to be tenable for Louisiana's tests. As an additional benefit, single year covariate models do not accentuate the lost records/linkages problems that arise from grade retention (which is a significant issue in Louisiana due to high rates of retention). Obviously, a student taking the 4th grade assessment in two consecutive years cannot be analyzed jointly with students who are taking tests at two different grade levels.

The tradeoffs between analytic power, bias reduction, measurement assumptions, and missing data issues among the alternative approaches to fitting longitudinal nested data within education remains an active area of investigation (Hong & Raudenbush, 2008; Lockwood & McCaffrey, 2007). However, the lack of available assessment instruments that are vertically aligned and univariate across grades limits their applicability to the data that are available in Louisiana. The current analyses replicate the prior work and adopted an HLM covariate approach to the data (Noell, 2006; Noell et al., 2007).

Scope and Timeliness of the Current Report

Two issues are noteworthy regarding this report that are distinct from the technical issues surrounding assembling, analyzing, and reporting the data. First, the report is being provided to the Board of Regents 2-3 months behind its projected release. The work was originally completed in August 2008. However, review of the numbers of graduates from various programs and the resulting programs whose results could be reported suggested problems with the reporting of program completers. Review of the data identified a number of reporting anomalies regarding who completed which programs and these were corrected. Additionally it was recognized that the program completer reporting mechanism that is used for the universities needed to be extended to private providers. As a result, updated and corrected data for TPP completers was provided to the research team in September. Subsequent to this, databases relevant to teachers and classes were rebuilt and analyses that included TPP effects were rerun. The need to repeat these substantial parts of the research work delayed release of the report.

The second major contextual issue surrounding this report has to do with its scope. Data are not reported for the majority of teacher preparation programs in Louisiana. This arises because the State's certification structure and all of its TPPs were redesigned in the period of 2000-2003. As a result, the majority of the new teachers captured in this report (school years 2004-2005 to 2006-2007) graduated from the programs that have since been retired. In order for the data to be informative regarding current programs, the report was limited to graduates of the ***current*** redesigned teacher preparation programs. As the stream of new teachers continues to shift from the original to the redesigned programs, data will become increasingly available that include most of Louisiana's TPPs. As a result of the impact of redesign, this report contains data for alternative pathway to certification programs only. This emerged because they completed redesign first and take less time than undergraduate paths to prepare teachers. Table 1 below was provided by Dr. Jeanne Burns and describes the transition from original to redesigned teacher preparation programs.

Table 1: *Percent of Program Completers in Pre-redesign and Post-redesign Programs by Year and Program Type*

Types of Programs	2002-2003	2003-2004	2004-2005	2005-2006
Undergraduate Programs (Pre-Redesign)	100%	100%	93%	75%
Undergraduate Programs (Post-Redesign)	0	0	7%	25%
Alternate Programs (Pre-Redesign)	75%	37%	24%	14%
Alternate Programs (Post Redesigned)	25%	63%	76%	86%

II. Data Merging Process

Data for the 2004-2005 and 2005-2006 academic years were merged following a process that has been described previously and was substantially replicated with the current year data (Noell, 2006; Noell et al., 2007). The data from the three individual school years were then combined to form a larger multiyear data set (described below) for the purpose of assessing TPPs.

Data for 2006-2007 were drawn from the standardized test files (*i*LEAP and *LEAP-21*) for spring 2006 and 2007, the curriculum database linking students to teachers, and supplemental student databases. The testing and supplemental databases provided data regarding attendance, enrollment, disability status, free lunch status, and demographic variables (e.g., race and gender). Data regarding teachers was drawn from the certification database, teacher attendance, and teacher demographic data obtained from the Louisiana Department of Education. Additionally, all TPP completers were identified through data provided to the Board of Regents by the TPPs. A multistage process was used to create longitudinal records for students describing achievement, attendance, and demographic factors across years. Similarly, teacher data were merged to create complete records for preperation, attendance, and certification. The student and teacher databases were then linked through the curriculum database.

Initial work was undertaken to resolve duplicate records and multiple partially complete records that described the same student within the separate databases. Following this work, data files were merged in a series of steps and a further round of duplication resolution was undertaken. Students' data were linked across years based upon unique matches on multiple identifiers used in each stage of the matching process. Student records that remained unmatched were then examined for a potential unique match through a layered series of comparisons. The matching process included six stages that were implemented hierarchically and that required unique matches on at least three identifying variables in order for a match to be established. Additional details of this process are available from the first author.

Table 2 describes the number of records available and the percentage of the total records that were matched at that stage. Mathematics and science are provided as examples of the merging process as language arts is similar to mathematics and social studies is similar to science. The difference between these clusters is the result of an assessment in 9^{th} grade in mathematics and language arts, but not science and social studies.

Several important decision points are noteworthy. Initial records were limited to students who completed an assessment in grades 4-9 to permit the availability of one year prior achievement data. The percentage of students whose 2007 test records could be matched to 2006 test records was high, but somewhat less than previous years. This modest attenuation is attributable to the disruptive effects of Hurricanes Katrina and Rita in the 2005-2006 reducing the available assessment records for that school year. The records available for analysis were further attenuated by the number of students whose matched data were not consecutive grades (e.g., 3^{rd} to 4^{th}). Some students were retained in grade and a surprisingly large number of students were promoted two grades in a single year. Obviously the meaning of taking the same test two years in a row or completing assessments separated by more than one grade level differs from taking tests in the expected sequence. As a result they were excluded from analyses. Additionally, in order to be included in the analyses, the student was required to be enrolled in the same school from September 15, 2006 until March 15, 2007. Because the student-teacher-course nexus data are only collected once per year, once a student moves it is not possible to ascribe subsequent instruction to a particular teacher. Finally, in order to be included in the analyses, the students' attendance and achievement records had to be matched to the curriculum data to identify which courses the students took and who taught those courses.

Table 2: *Cases Available at Each Stage of the Matching Process*

	Mathematics	**Science**
Assessed students grades 4-9	337,093	269,196
Matched to 2006 data	304,540 (90.3%)	249,664 (92.7%)
Consecutive grades assessed	257,132 (76.3%)	217,200 (80.7%)
Single primary school of attendance In curriculum database	243,532 (72.2%)	204,834 (76.1%)

Note. The percentage in parentheses within each cell is the percentage of the total records available for analysis at that stage of database construction.

Once students' achievement, demographic, attendance, and course enrollment records were linked, these data were linked to information about their teachers. This included teacher certification data obtained from the Louisiana Department of Education's Division of Planning, Analysis, and Information Resources and preparation data obtained from the Louisiana Board of Regents. Course codes were collapsed into groups that were associated with specific test areas (i.e., mathematics, reading, language arts, science, and social studies). For example, 4^{th} grade reading was associated with reading tests and Life Science with science tests. Course codes that could not reasonably be linked to a standardized test (e.g., Jazz Ensemble) were dropped. Students who had more than one teacher in a content area were included for each teacher, but their weight was reduced in proportion to the number of classes in that content area in which the student was enrolled. So for example, if a student was enrolled in two mathematics classes that student would have a record linked to each mathematics teacher, but each was weighted 0.5.

III. Preliminary Analyses

Prior to analyses that linked students and teachers within an HLM, a series of statewide ordinary least squares (OLS) regression analyses were conducted to examine general patterns in the data. Preliminary regression analyses for the 2005 and 2006 data have been reported previously. In order to assure that results of the OLS analysis are directly relevant to the subsequent HLM analyses, regression analyses were conducted on those student test records that were included in the HLM analyses. Test scores were standardized to a mean of zero and unit standard deviation within grade and year. Variables were entered sequentially in blocks to examine the predictive power of conceptually meaningful blocks of variables: prior achievement, demographic factors, and attendance. Results for all content areas are presented below.

Table 3: *Mathematics Statewide Regression Analyses for 2007*

Predictors	Multiple Correlation
Z-score Prior Year Math	.807
Z-scores Prior Year Achievement	.823
Z-scores Prior Year Achievement Student demographic factors	.829
Z-scores Prior Year Achievement Demographic & attendance	.830

Table Note. *n* = 243,532, *Prior Year achievement* includes the *Z*-scores for reading, language arts, mathematics, science, and social studies. *Student demographic factors* included were free lunch status, reduced price lunch, gifted status, primary special education diagnosis (codes for emotionally disturbed, specific learning disability, mild mental retardation, other health impaired, speech/language concerns, and other special education diagnosis), limited English proficiency status, gender, Section 504 eligibility, and minority status (codes for Asian American, African American, Hispanic, and Native American). *Attendance* was the number of days the student was absent.

Table 4: *Reading Statewide Regression Analyses for 2007*

Predictors	Multiple Correlation
Z-score Prior Year Reading	.709
Z-scores Prior Year Achievement	.764
Z-scores Prior Year Achievement Student demographic factors	.770
Z-scores Prior Year Achievement Demographic & attendance	.771

Table Note. *n* = 168,814. All variables were entered as in Table 3, see the note above.

Table 5: *Language Arts/Writing Statewide Regression Analyses for 2007*

Predictors	Multiple Correlation
Z-score Prior Year Writing	.711
Z-scores Prior Year Achievement	.749
Z-scores Prior Year Achievement Student demographic factors	.765
Z-scores Prior Year Achievement Demographic & attendance	.768

Table Note. *n* = 243,465. All variables were entered as in Table 3, see the note above.

Table 6: *Science Statewide Regression Analyses for 2007*

Predictors	Multiple Correlation
Z-score Prior Year Science	.755
Z-scores Prior Year Achievement	.803
Z-scores Prior Year Achievement Student demographic factors	.810
Z-scores Prior Year Achievement Demographic & attendance	.811

Table Note. *n* = 204,834. All variables were entered as in Table 3, see the note above.

Table 7*: Social Studies Statewide Regression Analyses for 2007*

Predictors	Multiple Correlation
Z-score Prior Year Science	.713
Z-scores Prior Year Achievement	.775
Z-scores Prior Year Achievement Student demographic factors	.780
Z-scores Prior Year Achievement Demographic & attendance	.782

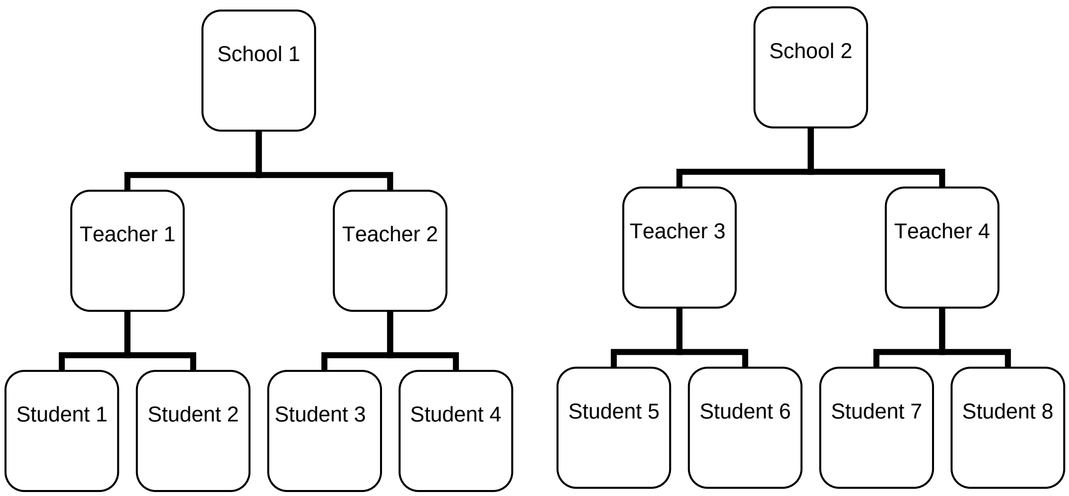
Table Note. n = 203,400. All variables were entered as in Table 3, see the note above.

Across all content areas, prior year academic achievement was sufficient to obtain a reasonably strong prediction of current year achievement and accounted for the majority of the multiple correlation (*r* .71 to .81). Each subsequent block of predictors accounted for a small, but statistically significant, additional portion of the variance in current year achievement. In all cases, the increment in multiple *r* and as a result shared variance dropped substantially as the blocks of predictors moved from achievement to demographic factors to attendance. The final multiple correlations ranged from .77 for Language Arts to .83 for Mathematics.

IV. Building the Base Model of Student Achievement Prior to VAA

Replicating the approach used in Noell (2006) and Noell et al. (2007), the educational assessment data were analyzed using hierarchical linear models (HLM; McCulloch & Searle, 2001; Raudenbush & Bryk, 2002). Hierarchical models were developed with students nested within classrooms that were in turn nested within schools. Interested readers may choose to consult Noell et al. for a detailed discussion of the variance apportionment between levels of the model, alternative models, and the impact of using a covariate adjustment approach to modeling results. This information will not be repeated here. Figure 1 below depicts the nesting structure that was employed.

Figure 1: *Nesting Structure of Students with Teachers and Teachers within Schools*



Building the current models. The modeling approach was somewhat parallel to Tekwe and colleagues (2004) in general strategy and followed by the VAA-TPP in the previous two years. Model development was completed independently for each school year 2004-2005, 2005-2006, and 2006-2007. The results of model development for 2004-2005 and 2005-2006 have been described previously (Noell et al., 2007). This report will focus on the results of model development for the 2007 assessments. The approach was replicated across mathematics, reading, language arts, science, and social studies. Error at each of the three levels (student, teacher, and school) was assumed to be normally distributed with a mean of 0 and common variance at that level. An initial 3 level model was specified in which achievement was modeled with no prior predictors as a basis for comparison with more complex models. Students' prior achievement in language arts, mathematics, reading, science, and social studies were entered as a block as fixed effects. All effects were significant in all content areas and were retained. Next, the 16 demographic variables employed in the regression analyses described above and student absences were entered as a block. Variables were then removed one at a time in order of the lowest t value until all remaining effects were significant at $p < .01$.

The decision to include student absences in the model will be evaluated as problematic by some readers. Some teachers will influence the level of student absences by the manner in which they teach and interact with students. This can result in higher or lower levels of absence. However, given that the students contributing to the analyses are minors typically between 8 and 15 years-of-age, their choice in whether or not to attend school will typically be strongly bounded by parental intervention. This is not so much an issue of absolute contributions but relative contribution to student absence. The authors adopted the assumption that students' absences were likely to be determined to a greater extent by variables that are beyond teacher control such as illness, parental choice, and chronic truancy than they are by student-teacher interaction. As a result student absences were retained as a potential predictor of student achievement.

Once a model for student level achievement was developed, several classroom variables were examined. These variables were entered at the teacher/classroom level and were conceptualized as contextual factors that may moderate student achievement in addition to teachers. The variables that were examined are presented in Table 8.

Table 8: *Classroom Level Variables*

Variable
Percentage of students who were male
Percentage of students who were minorities
Percentage of students who received free lunch
Percentage of students who received reduced price lunch
Percentage of students who were in special education
Percentage of students who were identified as gifted
Percentage of students who exhibited limited English proficiency
Class mean prior achievement in ELA
Class mean prior achievement in mathematics
Class mean prior achievement in science
Class mean prior achievement in social studies
Teacher absences

As with the student level demographic factors these classroom variables were entered as a block and removed one at a time in order of smallest *t* value for the coefficient. Once all effects were significant at the .01 level, the model for that content area was finalized. The same modeling process was then implemented across content areas for level 3 of the model (schools). The variables that were initially entered as a block are listed in Table 9.

Table 9: *School Level Variables*

Variable
Percentage of students who were male
Percentage of students who were minorities
Percentage of students who received free lunch
Percentage of students who received reduced price lunch
Percentage of students who were in special education
Percentage of students who were identified as gifted
Percentage of students who exhibited limited English proficiency
Percentage of students identified as protected by Section 504
Class mean prior achievement in ELA
Class mean prior achievement in mathematics
Class mean prior achievement in science
Class mean prior achievement in social studies
Percentage of students reported disrupted by hurricane

The following tables present the variables that were retained at the student, teacher, and school levels for each content area prior to consideration of teacher preparation effects. In all cases models were developed for intercepts as outcomes. At level 1 (students), prior achievement, demographic variables, and attendance were retained as predictors of test performance. At level 2, (teachers) classroom covariates were entered as predictors of the level 1 intercept (classroom mean) only and this effect was modeled as random. No classroom level predictors were entered for student level coefficients and student level coefficients were fixed. At level 3 (schools), school building level covariates were entered as predictors of the classroom intercept (school mean) only and this effect was modeled as random. No school building level predictors were entered for classroom level coefficients, and classroom level coefficients were fixed. These model specifications were adopted to enhance the interpretability of the data and were guided by the current research questions.

In summary, classroom and school building level covariates were used to adjust intercepts for students and classrooms respectively. No covariates were used to predict lower level coefficients and all coefficients were treated as fixed. Error variance was modeled for intercepts only. A simplified presentation of the model is provided below. Only equations for intercepts are presented. All other equations (e.g., the level 2 and level 3 models for level one coefficients) were modeled as fixed and not varying. In the equations presented below, $\sum$ is used to indicate summing across the p, q, and s coefficients at the student, teacher, and school levels of the model respectively.

Level 1: Students

$$Y_{ijk} = \pi_{0jk} + \sum(\pi_{pjk})a_{pijk} + e_{ijk}$$

where

Y_{ijk} is the achievement of student i in class j at school k in the target subject

π_{0jk} is the mean achievement for classroom j at school k

π_{pjk} are the *p* coefficients that weight the contribution of the student level data in the prediction of Y for $p = 1$ to the total number of coefficients

a_{pijk} are the student level data (prior achievement, demographic variables, and attendance) that predict achievement for $p = 1$ to the total number of data points

e_{ijk} the student level random effect, the deviation of the predicted score of student i in classroom j in school k from the obtained score

Level 2: Classrooms

$$\pi_{0jk} = \beta_{00k} + \sum(\beta_{q0k})X_{q0jk} + r_{0jk}$$

where

π_{0jk} is the mean achievement for classroom j at school k

β_{00k} is the mean achievement for school k

β_{q0k} are the q coefficients that weight the weight the relationship between the classroom characteristics and π_{0jk}, q = 1 to the total number of coefficients

X_{q0jk} are the classroom level data that are used to predict achievement; this is also the location in the model at which codes for recent TPP completers are entered (described below)

r_{0jk} the classroom level random effect, the deviation of classroom jk's measured classroom mean from its predicted mean

Level 3: Schools

$$\beta_{00k} = \gamma_{000} + \sum(\gamma_{s00})W_{s00k} + u_{00k}$$

where

β_{00k} is the mean achievement for school k

γ_{000} is the grand mean achievement in the target subject

γ_{s00} are the s coefficients that weight the weight the relationship between the school characteristics and β_{00k} for s = 1 to the total number of coefficients

W_{s00k} are the school level data that are used to predict achievement

u_{00k} the school level random effect, the deviation of school k's measured classroom mean from its predicted mean

The values presented in the tables below are the final values that were obtained prior to entering teacher preparation program codes into the model. The coefficients for university preparation programs are presented in the section regarding the VAA of teacher preparation.

Table 10: *Hierarchical Linear Model for Mathematics Achievement*

Model Level	Variables Entered	Coefficient	(CI)
Student level variables	Gender (Male)	1.5	(1.2, 1.7)
	African American	-5.3	(-5.6, -4.9)
	Asian American	5.3	(4.4, 6.3)
	Native American	-1.8	(-3.0, -0.6)
	Emotionally Disturbed	-4.8	(-7.1, -2.6)
	Speech and Language	-2.5	(-3.4, -1.6)
	Mild Mental Retardation	-12.1	(-14.4, -9.8)
	Specific Learning Disability	-6.6	(-7.4, -5.9)
	Other Health Impaired	-7.9	(-9.0, -6.8)
	Special Education - Other	-9.0	(-11.0, -7.1)
	Gifted	9.2	(8.4, 10.0)
	Section 504	-4.3	(-4.9, -3.7)
	Free Price Lunch	-1.7	(-1.9, -1.4)
	Reduced Price Lunch	-0.8	(-1.2, -0.5)
	Student Absences	-0.2	(-0.3, -0.2)
	Prior Year Math Test	26.8	(26.6, 27.0)
	Prior Year Reading Test	1.3	(1.1, 1.5)
	Prior Year Science Test	4.8	(4.6, 5.0)
	Prior Year Social Studies Test	2.5	(2.3, 2.8)
	Prior Year Language Arts Test	3.1	(2.9, 3.4)
Classroom variables	Teacher Absences	-0.1	(-0.1, 0.0)
	% Special Education	-0.7	(-1.0, -0.5)
	% Free Price Lunch	-1.4	(-1.7, -1.0)
	% Gifted	0.5	(0.2, 0.7)
	Mean Prior Year Math Test	-7.9	(-9.6, -6.2)
	Mean Prior Year Social Studies Test	5.7	(4.4, 7.1)
Building variables	Mean Prior Year Lang. Arts Test	-5.4	(-8.5, -2.2)
	Mean Prior Year Math Test	19.0	(15.1, 22.9)
	Mean Prior Year Science Test	-8.1	(-11.9, -4.4)
	% Minority	0.4	(0.1, 0.6)
	% Free Price Lunch	0.1	(0.0, 0.1)

The coefficients are scaled to the approximate standard deviation of the educational assessments (*i*LEAP and *LEAP*) used in Louisiana: 50. So after considering all other variables, a student who was identified as Emotionally Disturbed would be predicted to score 4.8 points lower than one who was not and a student who was gifted would be predicted to score 9.2 points higher in mathematics.

It is also important to recognize that the inclusion of teacher absences in the model will be regarded as problematic by some readers. To the extent that TPPs are more or less successful in preparing teachers who have poor or excellent work attendance this variable could be siphoning off some of the TPP effect. However, it may also be the case that factors beyond the control of universities are likely to be more determinative regarding teacher attendance. In particular teacher health and school district professional development requirements seem likely to have a larger impact on attendance than TPPs.

It is important to note that differences in how variables were scaled create the need for considerable caution in comparing the coefficients across different types of predictors. Demographic variables at the student level were coded 1 if present and 0 if absent. Prior achievement is measured in standard deviation units from the grand mean prior achievement. Classroom percentages are measured in 10% units, so that the value presented would be the expected change in students' scores if the percentage of the indicated group increased by ***10%***. Due to differences in scales of measurement and the meaning of the measurements it is difficult to make direct comparisons across different types of measures.

The largest single contributor to a student's mathematics achievement *among the achievement predictors* was his or her achievement in that domain the prior year. The coefficient for prior achievement in mathematics was more than five times the value of any other prior achievement variable's coefficient.

Among the demographic variables, the special education disabilities Mild Mental Retardation, Learning Disability, Other Health Impaired, and Special Education – Other had large negative coefficients. In contrast, giftedness had a large positive coefficient. Males had a slight advantage over females. The magnitude of the negative coefficient for a student being African American *after accounting for prior achievement, poverty, and disability* should be a concern to educators beyond the consideration of teacher preparation.

The magnitude of the coefficient for student absences may surprise some readers, however it is important to note that this is the effect for each day absent. So a student who was absent 20 days would be predicted to score 4 points lower than one with perfect attendance.

Classroom demographic variables loaded in what would be the commonly expected direction with the exception that mean prior achievement in mathematics loaded negatively. This is a phenomenon that has emerged across some content areas and was evident for mathematics in 2006. The most plausible interpretation appears to be that this serves as corrective factor for overly positive predictions for clusters of high performing students or overly negative predictions for clusters of poor performing students. It may represent a regression toward the mean. It is important to recall that this inverse phenomenon only occurs in a context saturated with considerable information at the student level. Prior work has shown that examined in isolation, high mean prior student achievement predicts higher current year achievement (Noell et al., 2007). Generally the magnitude of the classroom coefficients was not large when considered in regard to what was measured for each predictor.

The school building coefficients initially appear difficult to interpret given their large magnitude and mixture of expected (e.g., the positive coefficient for mean prior mathematics achievement) and paradoxical (e.g., the negative coefficient for prior language arts) coefficients. The magnitude of the prior achievement effects is likely the result of scaling. Prior achievement is measured in standard deviation units and having entire school buildings whose mean prior achievement is a standard deviation above the mean is a large and rare effect. This may have accentuated the magnitude of these coefficients substantially. As to the mixture of expected and paradoxical coefficients, these only emerge in a model saturated with a tremendous amount of information at the student and classroom level. It is likely that these coefficients are capturing small patterns in the residual variance to enhance the fit of the model are not directly interpretable. It is also worth noting that the prior work of the VAA-TPP has found the proportion of the variance that lies between schools to be quite small (approximately 3.3%).

Table 11: *Hierarchical Linear Model for Reading Achievement*

Model Level	Variables Entered	Coefficient	(CI)
Student level variables	Gender (Male)	-2.9	(-3.2, -2.5)
	African American	-3.7	(-4.2, -3.2)
	Limited English Proficiency	-3.6	(-5.2, -2.1)
	Mild Mental Retardation	-15.1	(-18.1, -12.0)
	Other Health Impaired	-10.1	(-11.7, -8.6)
	Specific Learning Disability	-12.9	(-14.1, -11.8)
	Special Education - Other	-6.8	(-9.5, -4.1)
	Speech and Language	-4.5	(-5.6, -3.4)
	Emotionally Disturbed	-4.0	(-7.2, -0.8)
	Gifted	7.6	(6.5, 8.6)
	Section 504	-6.2	(-7.1, -5.3)
	Free Price Lunch	-2.7	(-3.1, -2.3)
	Reduced Price Lunch	-0.8	(-1.3, -0.3)
	Student Absences	-0.1	(-0.1, -0.1)
	Prior Year Math Test	3.4	(3.1, 3.8)
	Prior Year Reading Test	14.7	(14.4, 15.1)
	Prior Year Science Test	9.2	(8.9, 9.6)
	Prior Year Social Studies Test	6.8	(6.5, 7.1)
	Prior Year Language Arts Test	4.7	(4.4, 5.0)
Classroom variables	Teacher Absences	-0.03	(-0.06, -0.02)
	% Special Education	-0.6	(-0.9, -0.4)
	% Free Price Lunch	-1.3	(-1.7, -1.0)
	% Gifted	0.4	(0.1, 0.6)
	% Minority	-0.4	(-0.9, 0.0)
	% Gender (Male)	-0.4	(-0.7, -0.1)
	Mean Prior Year Math Test	-5.9	(-7.3, -4.5)
Building Variables	% Minority	0.7	(0.3, 1.2)
	% Free Price Lunch	0.1	(0.0, 0.1)
	Mean Prior Year Reading Test	14.3	(10.5, 18.0)
	Mean Prior Year Science Test	-5.9	(-9.5, -2.4)

The pattern of coefficients for reading closely parallels the coefficients for mathematics with a few exceptions. The coefficient for males reversed in direction, suggesting a relative advantage for girls over boys in reading. Although prior achievement in reading was the best predictor of current year reading, the coefficients for prior achievement domains were less differentiated. The continued negative loading for African American students after inclusion of prior achievement and free lunch status should be a source of concern.

Table 12: *Hierarchical Linear Model for Language Arts Achievement*

Model Level	Variables Entered	Coefficient	(CI)
Student level variables	Gender (Male)	-11.3	(-11.6, -10.9)
	African American	2.0	(1.6, 2.4)
	Asian American	5.8	(4.6, 7.0)
	Hispanic American	2.3	(1.4, 3.3)
	Limited English Proficiency	-2.4	(-3.8, -1.1)
	Emotionally Disturbed	-7.5	(-10.4, -4.7)
	Speech and Language	-3.5	(-4.5, -2.5)
	Mild Mental Retardation	-20.3	(-23.6, -17.0)
	Other Health Impaired	-9.2	(-10.5, -8.0)
	Specific Learning Disability	-13.7	(-14.7, -12.7)
	Special Education - Other	-5.3	(-7.8, -2.7)
	Gifted	9.0	(8.1, 9.9)
	Section 504	-7.0	(-7.7, -6.2)
	Free Price Lunch	-2.0	(-2.3, -1.7)
	Reduced Price Lunch	-1.2	(-1.7, -0.7)
	Student Absences	-0.4	(-0.4, -0.3)
	Prior Year Math Test	7.5	(7.3, 7.8)
	Prior Year Reading Test	5.5	(5.2, 5.7)
	Prior Year Science Test	2.7	(2.4, 2.9)
	Prior Year Social Studies Test	4.0	(3.7, 4.3)
	Prior Year Language Arts Test	16.9	(16.5, 17.4)
Classroom variables	Teacher Absences	-0.04	(-0.06, -0.02)
	% Special Education	-0.7	(-0.9, -0.4)
	% Free Priced Lunch	-1.6	(-1.9, -1.2)
	% Gender (Male)	-0.5	(-0.9, -0.2)
	Mean Prior Year Math Test	-3.9	(-5.6, -2.1)
	Mean Prior Year Social Studies Test	1.6	(-0.1, 3.3)
Building variables	Mean Prior Year Language Arts Test	8.4	(6.5, 10.4)
	% Free Price Lunch	0.2	(0.1, 0.2)
	% Reduced Price Lunch	-1.5	(-2.6, -0.4)

The pattern of coefficients for language arts closely parallels the coefficients for reading with a few exceptions. One notable difference is that the negative coefficient for males is a great deal larger for language arts than it is in reading. Although the authors have no data to substantiate this hypothesis, we are curious if in the constructed response portion of the language arts test boys' responses were scored more poorly due to the poor legibility of handwriting that is more prevalent among boys. It is also interesting to note that language arts is the only content area in which the coefficient for African American students was positive.

Table 13: *Hierarchical Linear Model for Science Achievement*

Model Level	Variables Entered	Coefficient	(CI)
Student level variables	Gender (Male)	3.8	(3.5, 4.0)
	African American	-7.0	(-7.4, -6.6)
	Limited English Proficiency	-1.9	(-3.0, -0.8)
	Emotionally Disturbed	-5.5	(-8.1, -2.9)
	Mild Mental Retardation	-10.8	(-13.3, -8.2)
	Specific Learning Disability	-3.5	(-4.3, -2.7)
	Other Health Impaired	-4.8	(-5.9, -3.7)
	Special Education - Other	-10.7	(-12.9, -8.4)
	Gifted	6.2	(5.4, 7.0)
	Section 504	-1.5	(-2.2, -0.9)
	Free Price Lunch	-2.2	(-2.6, -1.9)
	Reduced Price Lunch	-0.8	(-1.2, -0.3)
	Student Absences	-0.2	(-0.2, -0.2)
	Prior Year Math Test	8.1	(7.9, 8.4)
	Prior Year Reading Test	7.9	(7.7, 8.2)
	Prior Year Science Test	13.9	(13.6, 14.2)
	Prior Year Social Studies Test	7.2	(6.9, 7.4)
	Prior Year Language Arts Test	1.3	(1.0, 1.5)
Classroom variables	Teacher Absences	-0.04	(-0.05, -0.02)
	% Special Education	-0.6	(-0.8, -0.3)
	% Free Price Lunch	-1.0	(-1.4, -0.7)
	% Gifted	0.7	(0.4, 1.0)
	% Minority	-0.3	(-0.5, -0.1)
	Mean Prior Year Math Test	-5.1	(-6.6, -3.7)
Building variables	% Free Price Lunch	0.1	(0.1, 0.2)
	Mean Prior Year Social Studies Test	8.8	(7.0, 10.6)

The base model for science achievement shares some features with both the mathematics model and the reading model. Similar to the results for mathematics, gender (male) loaded positively and being identified as an African American loaded negatively. Similar to reading, prior achievement in the content, science in this case, was the strongest predictor among the prior achievement variables, but results were not as starkly differentiated as they were in mathematics. As with the other content areas, prior mathematics achievement for the class loaded in the opposite of the expected direction in a model saturated with student level information.

Table 14: *Hierarchical Linear Model for Social Studies Achievement*

Model Level	Variables Entered	Coefficient	(CI)
Student level variables	Gender (Male)	3.0	(2.7, 3.3)
	Asian American	5.6	(4.5, 6.7)
	Hispanic American	3.2	(2.2, 4.2)
	African American	-1.7	(-2.1, -1.3)
	Section 504	-3.6	(-4.3, -2.9)
	Emotionally Disturbed	-5.3	(-7.9, -2.6)
	Mild Mental Retardation	-7.7	(-10.3, -5.0)
	Other Health Impaired	-6.5	(-7.7, -5.3)
	Specific Learning Disabilities	-5.1	(-5.9, -4.3)
	Special Education - Other	-5.9	(-8.2, -3.6)
	Gifted	8.5	(7.7, 9.3)
	Student Absences	-0.3	(-0.3, -0.2)
	Free Price Lunch	-3.1	(-3.4, -2.7)
	Reduced Price Lunch	-1.5	(-2.0, -1.0)
	Prior Year Math Test	4.3	(4.0, 4.5)
	Prior Year Reading Test	8.6	(8.3, 8.9)
	Prior Year Science Test	9.6	(9.3, 9.8)
	Prior Year Social Studies Test	12.8	(12.5, 13.1)
	Prior Year Language Arts Test	2.2	(1.9, 2.4)
Classroom variables	% Free Price Lunch	-1.1	(-1.4, -0.7)
	Teacher Absences	-0.1	(-0.1, 0.0)
Building variables	Mean Prior Year Reading Test	-6.4	(-10.3, -2.5)
	Mean Prior Year Science Test	-9.0	(-13.8, -4.3)
	Mean Prior Year Social Studies Test	22.0	(17.5, 26.5)
	% Limited English Proficiency	1.4	(0.4, 2.4)
	% Free Price Lunch	0.1	(0.0, 0.1)

As with of the other content areas, prior achievement in the domain was the strongest predictor of current year social studies achievement. Similar to mathematics large coefficients emerged at the school level whose directions were mixed, including effects that would be in the expected direction (social studies) and ones that would be in the opposite of the expected direction in models lacking student level predictors (e.g., science). As with all of the content areas, all disability status variables loaded negatively and being gifted was advantageous. In social studies the coefficient for males was positive and the coefficient for African American students was negative.

Summary. Generally, the student level models had much in common across content areas. For all areas, prior achievement in the target content area had the largest coefficient *among prior achievement variables*, with achievement in the other four content areas loading to various degrees. Having a special education diagnosis was a consistent, strong negative predictor of achievement and in many cases (e.g., Mild Mental Retardation) the effect was large. Student absences and free lunch status exhibited consistent relatively small coefficients. Among the ethnicity factors, no single variable was consistently statistically significant and always loaded in the same direction. However, status as an African American loaded in all of the models and loaded negatively in four of five models.

The only completely consistent finding at the classroom level was the small negative loading for teacher absences. Beyond that, coefficients for non-achievement demographic variables at the classroom level exhibited small positive coefficients for increasing percentages of advantaged groups (e.g., gifted students) and negative coefficients for disadvantaged students (e.g., disabled students). In four of five content areas, a counter intuitive loading occurred in which classroom mean prior achievement in mathematics loaded negatively. However, this is a phenomenon that only emerges in a model that is saturated with a tremendous amount of information about student achievement and demographic factors. The simple relationships are in the expected directions.

The variables that were statistically significant, the direction of the loadings, and the magnitude of those loading at the school building level might be best described as idiosyncratic across content areas. This finding may not be surprising when one considers the small amount of variance accounted for by the school building level of the model and the tremendous amount of information provided at the student and classroom level.

VI. Assignment of Teachers to Groups

The operational definition of "new teachers" that was employed in the prior VAA-TPP work of teachers in their first two years of teaching was carried forward in this year. The figures below present the mean coefficients across the three academic years that were extracted for the years of experience effect using teachers with more than 20 years experience as the comparative group. These analyses were based on the years experience variable in the teacher certification databases provided by the Louisiana Department of Education. This variable has some issues regarding its interpretability (see Noell et al., 2007 for an explanation).

Figure 2

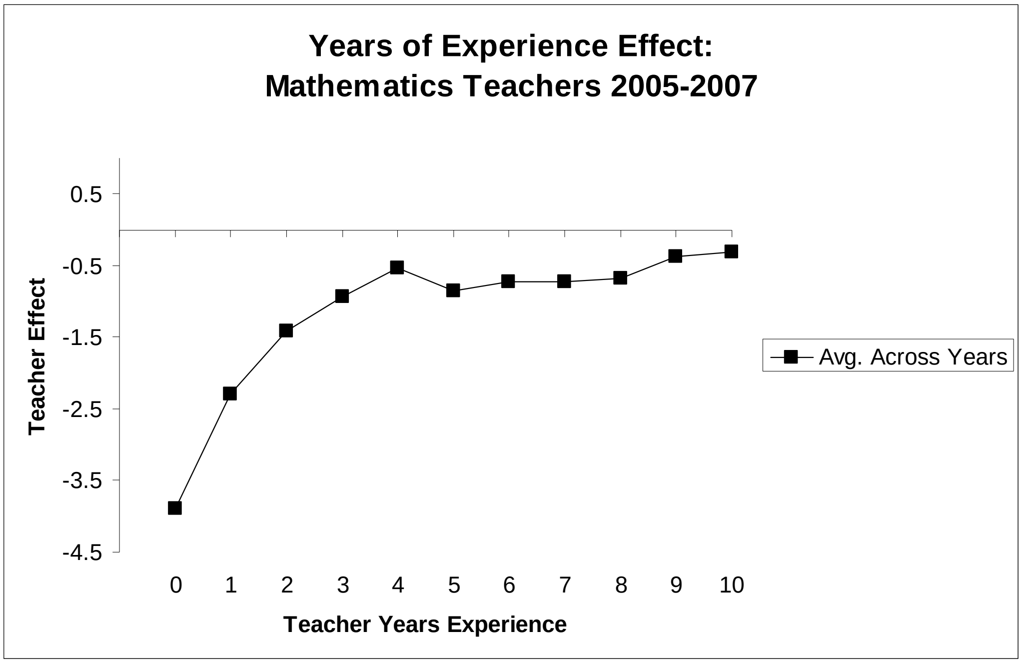


Figure note. The zero line on this graph represents the average effect for teachers with 21-30 years experience.

Figure 3

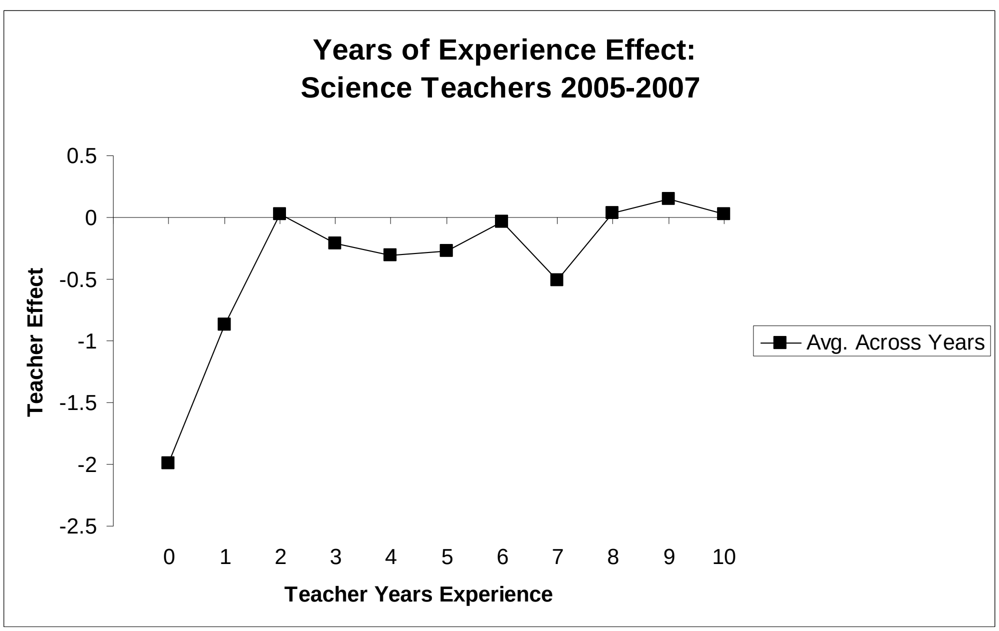


Figure note. The zero line on this graph represents the average effect for teachers with 21-30 years experience.

Figure 4

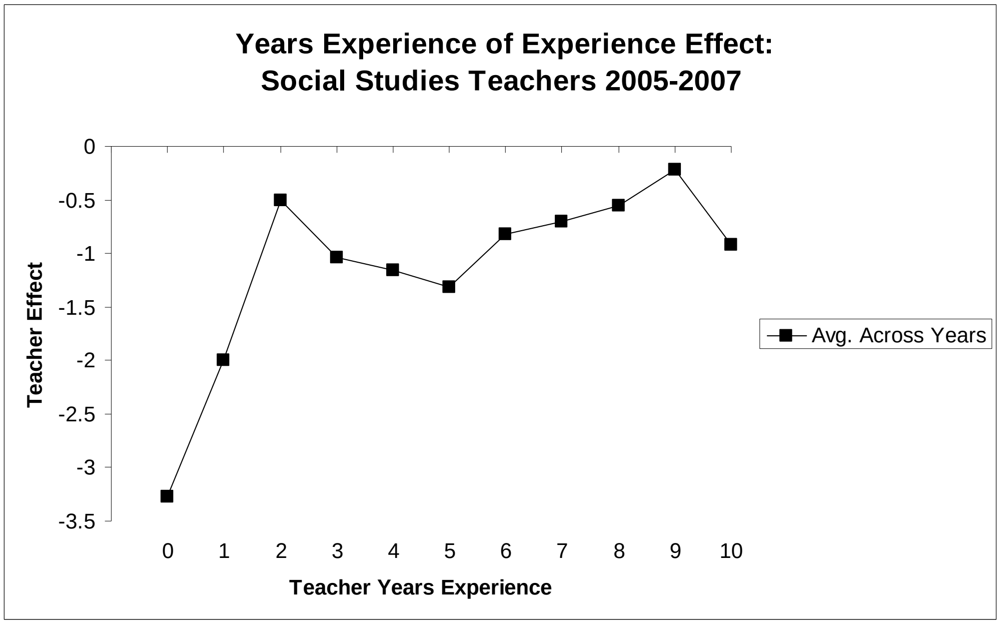


Figure note. The zero line on this graph represents the average effect for teachers with 21-30 years experience.

Figure 5

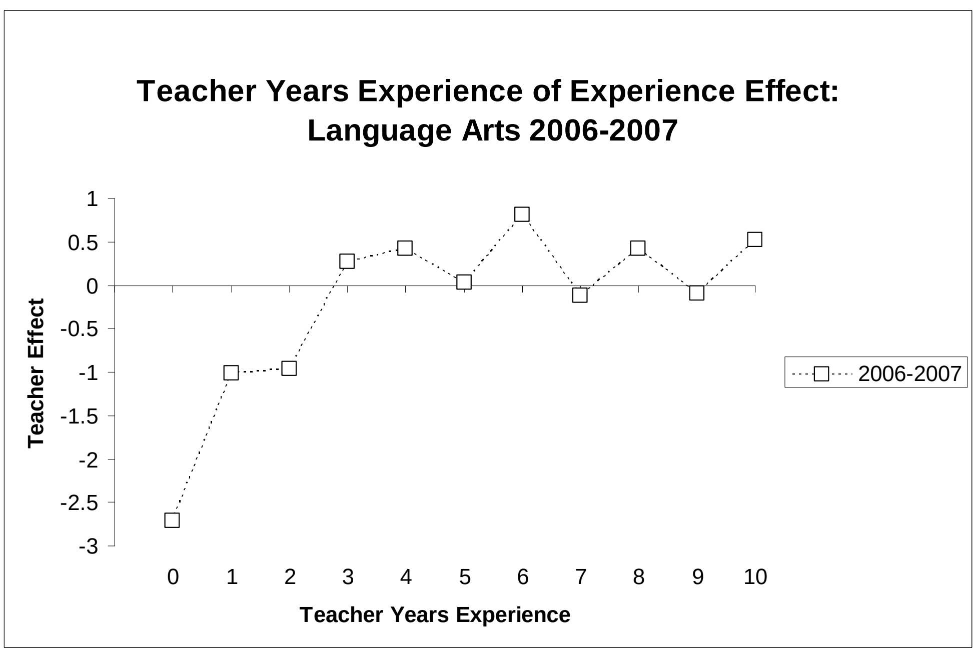


Figure note. The zero line on this graph represents the average effect for teachers with 21-30 years experience.

Based on the data and a desire to maintain a consistent operational definition of new teachers the following operational definition was maintained from the VAA-TTP report of 2007.

Table 15: *Teacher Group Assignment*

Group	Criteria
New teachers	1. Teachers in their first or second year of teaching after completing a teacher preparation program leading to initial certification. 2. Certified to teach in the content area. 3. Completed teacher preparation program within 5 years of starting teaching.
Regularly Certified Teachers	1. All other teachers holding a standard certificate. 2. Certified to teach in the content area assessed.
Other	1. Does not conform to any of the categories above.

All subsequent analyses were based upon this categorization combined with the teachers' preparation program that could lead to teacher certification.

VII. VAA of Teacher Preparation

Once the final models for student achievement nested within classrooms and schools were developed, these models were used to assess deviations in students' achievement that were associated with being taught by a new teacher from a particular teacher preparation program. This step was the VAA. TPP modeled at the teacher level by a series of codes representing being a new program completer from a particular preparation program. Each type of program was modeled separately for each provider: undergraduate, practitioner, master's degree, and non-master's certification only.

The coefficients for recent graduates of particular programs were modeled on the scale of the current *i*LEAP and *LEAP-21* tests due to their importance in high stakes assessment for promotion in grades 4 and 8 as well as their disproportionate weight in School Performance Scores. The tests for 2006 and 2007 had a mean of approximately 300 and a standard deviation of approximately 50 across content areas and grade levels. The results reported below are the mean expected effect for that teacher preparation program in comparison to experienced certified teachers.

Prior work (see Noell et al., 2007), used an examination of the ratio of variance within programs to variance between programs to arrive at the rule of only reporting results when data were available for 25 teacher within year observations of new teachers from a specific program. That rule was carried forward herein.

Combining Data Across Years

Following the analytic strategy developed in the VAA-TPP 2007 report, the three consecutive school years were jointly analyzed. The dependent variable was the target achievement test score. The predictor variables were those variables that were identified during model development for that year. All predictor variables for other years were set to 0 (interacted with year). The codes TPPs were not interacted with years allowing extraction of cross year coefficients and standard errors from the pooled data.

Additionally, teachers and schools were modeled independently across years. This specification has both analytic and pragmatic advantages. The analytic advantage of specifying schools as independent across years is that it avoids the problematic assumption that schools are the same organizational units across years. This is obviously not the case when schools are redistricted, have substantial changes in staff, or have their grade configuration revised. One disadvantage is that the model did not capitalize on the repeated observation of teachers across years. However, no software could be identified at the time that these analyses were completed that would allow for such a complex cross classification structure at the teacher level and that could also resolve a model with so many variables, individuals, and levels. As a result, a model was adopted that treated schools, teachers, and students as independent observations across years. Scientific Software is preparing software that would be able to solve such a complex model and the authors will be examining its application to this problem in the future.

Propensity Score Matching

An additional step was taken to reduce the likelihood that a relatively distinct pattern of new teacher placement would distort the TPP coefficients. Propensity score matching (PSM) was used to select a comparative group of teachers whose class compositions were similar in likelihood to the classes taught by new teachers from each TPP. A particular challenge to PSM in the current context is the use of multilevel models. In the current context, using a 1-to-n strategy for matching nearest neighbors with an *n* of 1, 5, or even 10 will lead to many or most teachers being the only teacher nested within their school. In these circumstances, estimation of the school level of the model will be poor because most of the information about the teachers in the school will be lost. This can lead to relatively severe distortion of models. Models using weighting can have similar consequences. In an HLM context, PSM has the paradoxical risk of introducing bias due to poor estimation of school effects.

To minimize this risk, an approach to PSM was adopted that was designed to maintain as many of the reasonable matches as was possible. The approach used herein was derived from the procedures described by Rubin and Thomas (2000). Rubin and Thomas describe using the nearest remaining neighbor matching strategy based upon logistic propensity scores using 1 to 1 and 1 to 5 matches. In an alternative they describe using a “coarse” (p. 574) caliper .2 to specify a range within which to perform Mahalanobis distance matching. Given the number of programs to match across multiple content areas, the thousands of included teachers, and the desire to maintain a large *n* when propensity matching permitted maintaining a reasonable three level model, an adaptation of these two procedures was used. In the current application a fine caliper that was 5% of the width of Rubin and Thomas’ coarse caliper was used (i.e., .01). However, all matches within a .01 caliper of any classroom from that specific TPP were retained for analysis.

For each TPP, a logistic regression was used to obtain the probability of a classroom being consistent with the classrooms of the new teachers from the TPP. This probability was then converted to a propensity score and all matches within plus or minus a caliper .01 for any classroom taught by a graduate of the TPP was selected for the comparison. HLM VAA analyses were conducted using both a statewide all teachers, students, and schools model and the sample of teachers obtained from the PSM. Aside from the sample difference, the models used in the whole sample and PSM models differed only modestly. In the PSM case, only the effect of the TPP of interest was included among the possible TPP effects. Also, in many cases variables had to be excluded in the PSM derived analyses as they resulted in singularities as no student in the PSM sample exhibited that characteristic. For example, some TPPs had no students with emotional disturbance in their classes and none were in the PSM sample. As a result this predictor was excluded. The comparison of the PSM and statewide whole sample models suggest they produce functionally the same result. The correlations for mathematics, reading, language arts, science, and social studies between the PSM coefficients and the coefficients for the full data set were r = .99, .99, .98, .99, and .99 respectively. Given the tremendous computational and data management burden created by the PSM process and the absence of evidence that it produces a different result than the whole sample approach, the whole sample data are presented below. The research team anticipates providing the whole sample data going forward.

Performance Bands for Mathematics, Science, and Social Studies

For the 2007 VAA-TPP report a series of five performance bands was developed in consultation with the then Commissioner of Higher Education and the Associate Commissioner for Teacher Education Initiatives. These levels were designed to create bands of performance that have some intuitive meaning and may help focus readers on clusters of performance rather than a continuous ranking in which the ordering between near neighbors is much more likely to be the result of measurement error than a meaningful difference. The performance levels are defined below.

Level 1 – Programs whose effect estimate is above the mean effect for experienced teachers by its standard error of measurement or more. These are programs for which there is evidence that new teachers are more effective than experienced teachers, but this is not a statistically significant difference. The difference between these programs and the mean for new teachers would commonly be statistically significant.

Level 2 – Programs whose effect estimate is above the mean effect for new teachers by its standard error of measurement or more. These are programs whose effect is more similar to experienced teachers than new teachers.

Level 3 – Programs whose effect estimate is within a standard error of measurement of the mean effect for new teachers. These are programs whose effect is typical of new teachers.

Level 4 – Programs whose effect estimate is below the mean effect for new teachers by its standard error of measurement or more. These are programs for which there is evidence that new teachers are less effective than average new teachers, but the difference is not statistically significant.

Level 5 – Programs whose effect estimate is statistically significantly below the mean for new teachers.

Tables 16-20 below present the VAA estimates for mathematics, reading, language arts, science, and social studies. The more liberal 68% CI was adopted for this report based on the assumption that for a formative assessment such as this, the consequences of false negatives, failing to identify an exemplary program or one that is struggling, are typically at least comparable to the risks of a false negative.

Table 16: *Teacher Preparation Program Coefficient for Mathematics*

Level	**Teacher Preparation Program**	**2005-2007 Estimate** (CI)	**Teachers**
1	New Teacher Project Practitioner TPP	3.1 (1.5, 4.7)	55
2	University of Louisiana - Monroe Master's Alt. Cert.	1.1 (-0.4, 2.6)	30
2	Northwestern State University Practitioner TPP	0.8 (-1.4, 3.0)	63
3	Louisiana College Practitioner TPP	-2.7 (-4.8, -0.6)	41
3	University of Louisiana - Lafayette NM/CO	-2.9 (-4.5, -1.2)	34
3	Louisiana Resource Center for Educators Practitioner TPP	-3.2 (-4.8, -1.6)	49

Note. The top number in the estimate cells is the mean adjustment to student outcome that would be expected based upon a standard deviation of 50. The numbers in parentheses are the 68% confidence intervals. The mean new teacher effect was -2.7.

Table 17: Teacher Preparation Program Coefficient for Reading

Level	**Teacher Preparation Program**	**2005-2007 Estimate** (CI)	**Teachers**
1	New Teacher Project Practitioner TPP	2.2 (0.7, 3.7)	41
1	Louisiana College Practitioner TPP	2.1 (0.3, 3.9)	35
2	Northwestern State University Practitioner TPP	0.6 (-1.0, 2.1)	53
3	University of Louisiana - Lafayette NM/CO	-2.4 (-4.2, -0.6)	41
5	Louisiana Resource Center for Educators Practitioner TPP	-6.2 (-8.2, -4.2)	35

Note. The top number in the estimate cells is the mean adjustment to student outcome that would be expected based upon a standard deviation of 50. The numbers in parentheses are the 68% confidence intervals. The mean new teacher effect was -1.8.

Table 18: Teacher Preparation Program Coefficient for Language Arts

Level	**Teacher Preparation Program**	**2005-2007 Estimate** (CI)	**Teachers**
1	University of Louisiana - Monroe Master's Alt. Cert.	2.7 (0.3, 5.0)	28
1	New Teacher Project Practitioner TPP	1.6 (0.2, 2.9)	56
2	Louisiana College Practitioner TPP	1.5 (-0.7, 3.6)	34
2	Northwestern State University Practitioner TPP	0.5 (-1.1, 2.1)	55
2	Nicholls State University Practitioner TPP	-0.3 (-1.7, 1.0)	27
3	Louisiana Resource Center for Educators Practitioner TPP	-1.8 (-3.3, -0.3)	42
4	University of Louisiana - Lafayette NM/CO	-4.6 (-6.7, -2.5)	43

Note. The top number in the estimate cells is the mean adjustment to student outcome that would be expected based upon a standard deviation of 50. The numbers in parentheses are the 68% confidence intervals. The mean new teacher effect was -1.8.

Table 19: *Teacher Preparation Program Coefficient for Science*

Level	**Teacher Preparation Program**	**2005-2007 Estimate** (CI)	**Teachers**
1	Northwestern State University Practitioner TPP	2.7 (1.5, 4.0)	50
1	University of Louisiana - Monroe Master's Alt. Cert.	1.7 (0.6, 2.8)	27
2	New Teacher Project Practitioner TPP	0.7 (-1.1, 2.4)	50
3	Louisiana College Practitioner TPP	0.4 (-2.3, 3.1)	33
3	University of Louisiana - Lafayette NM/CO	-0.9 (-3.0, 1.1)	29
3	Louisiana Resource Center for Educators Practitioner TPP	-1.3 (-2.7, 0.1)	35

Note. The top number in the estimate cells is the mean adjustment to student outcome that would be expected based upon a standard deviation of 50. The numbers in parentheses are the 68% confidence intervals. The mean new teacher effect was -1.1.

Table 20: *Teacher Preparation Program Coefficient for Social Studies*

Level	**Teacher Preparation Program**	**2005-2007 Estimate** (CI)	**Teachers**
1	University of Louisiana - Monroe Master's Alt. Cert.	2.8 (0.5, 5.0)	25
2	Louisiana College Practitioner TPP	2.6 (-0.5, 5.7)	39
2	Northwestern State University Practitioner TPP	0.8 (-0.4, 2.0)	46
3	New Teacher Project Practitioner TPP	-0.1 (-2.2, 2.1)	48
3	University of Louisiana - Lafayette NM/CO	-1.1 (-4.0, 1.8)	33
3	Louisiana Resource Center for Educators Practitioner TPP	-3.2 (-5.5, -0.9)	26

Note. The top number in the estimate cells is the mean adjustment to student outcome that would be expected based upon a standard deviation of 50. The numbers in parentheses are the 68% confidence intervals. The mean new teacher effect was -2.1.

Summary

Some interesting consistencies emerged in examining the results across the alternative TPPs represented in this year's data. The New Teacher Project, the Master's Program at the University of Louisiana at Monroe, and the Northwestern State University Practitioner Preparation programs had generally positive results. These programs exhibited consistent performance at Level 1 or Level 2 across content areas with only one Level 3 performance among them (NTP social studies). All three of these programs are producing teachers who in aggregate appear to be making a positive contribution to student achievement from the point of entering the classroom.

In contrast, the results for the Non-Master's Certification Only program at the University of Louisiana at Lafayette and the Practitioner Teacher Program at the Louisiana Resource Center for Educators were less positive. These programs consistently were the two programs with the most negative effect estimates among the programs represented in these data. So as not to paint an overly dire picture, it is important to recognize that in most cases their effect estimate fell in the average range for new teachers. For each of these two programs they had only one instance in which their program estimate fell below Level 3.

It is also interesting to note the degree of change and consistency in the estimates across years. At the outset it is important to acknowledge that using three year aggregate data will have a strong stabilizing effect on the data. Shifts in estimates ranged from range from 0 to 2.9 points in absolute value and from -2.9 to + 1.0 in magnitude. With the exception of the estimates for Louisiana College's Practitioner TPP all of the programs remained in the same performance level. Louisiana College Practitioner TPP estimates dropped one level across all content areas for those that were available in both 2007 and 2008. This may be an artifact of the small number of graduates the initial estimate was based on and a shift toward a more stable estimate or it may reflect a weaker cohort of new graduates for the 2006-2007 academic year.

VIII. Additional Detailed Family Demographic Variables

One concern that has arisen regarding the use of student achievement data to assess TPP efficacy is that it will not consider one of the major determinants of student achievement that is beyond the control of schools: families. This reasoned concern emerges from the studies showing that student achievement is correlated with family demographic factors, parenting practices, and family resources (Chatterji, 2006; Coleman, 1989; Downer & Pianta, 2006; Hill & Craft, 2003; White, 1982). Alternative arguments have been advanced that the strongest determinants of educational attainment are found in schools and/or that the influence of family factors is evident in prior educational attainment providing a reasonable means of estimating the impact of current educational inputs (Ballou et al., 2004; Sanders & Horn, 1998).

It may be that neither argument is completely correct, but that they are different possible outcomes that can occur depending on the available data systems. Specifically, it may be the case that when educational data are sparse, assessments are weakly related to one another, or that supplementary data such as attendance and special education disability status are not available that family data would substantially improve the prediction of student achievement. Alternatively, it may be the case that extensive educational data systems using measures that are strongly related to the outcomes of interest may account for so much of the finite shared explainable variance that the addition of family demographic variables would not improve prediction.

The issue of the adequacy of the educational data systems and the need for family demographic data is a practical rather than a theoretical issue for Louisiana's efforts to assess TPP. Beginning with the 2007 VAA-TPP report, the research team included data from an additional data collection of family demographic and family school interaction data to examine the extent to which these data would improve on the predictions obtained from the Louisiana's educational databases. The 2007 study found that the addition of the family variables accounted for an exceedingly small amount of additional variance in English Language Arts and mathematics (.002 to .005). It is also worth noting that the possibility for accounting for additional variance is limited given the large amount of variance accounted for by the educational data. The data collection procedure for the study of family demographic variables reported here in was the same as that described in the 2007 VAA-TPP report.The current year study obtained a sample that is 78% larger than the 2007 study and extended the analysis to all five content areas examined in this year's research: mathematics, reading, language arts, science, and social studies.

Participants. A stratified random sample of schools with students in grades 4-9 was identified and recruited to participate in the family survey data collection. The sample was stratified such that nine schools were recruited to represent each of three segments of the demographic distribution of schools in Louisiana. The segments were low (1st – 25th percentile), middle (26th – 75th percentile), and high (76th -99th percentile) on the variables of interest. Nine schools were randomly selected that represented the low, middle, and high end of the distribution of each variable. This sampling plan was intentionally mildly skewed to the tails of the distribution to increase the probability that the sample included sufficient representation of the range of variability evident in Louisiana, rather than being purely representative.

The variables that were used to stratify the sample were the percentage of students who were ethnic minorities, percentage of students receiving free or reduced price lunch, and the school performance score (SPS). The SPS is Louisiana's school accountability index that is comprised of weights of several variables, but the weighting is strongly dominated by achievement test data. Schools with high SPS will have students who are performing relatively well on the state assessments. A final additional demographic variable that was used for stratification was the locale code from the U.S. census for the school's zip code. Three strata were identified: urban (mid-size city code), suburban (two different urban fringe codes), and rural. Schools were offered $1500 to participate in data collection. Several schools that were contacted declined and replacement schools were identified through random selection.

Method. Schools that agreed to participate distributed survey packets to families by sending them home with students. The packet contained a cover letter soliciting participation and providing informed consent information, a brief survey (one page), and a sealable envelope in which to return the survey to the school. Parents were asked to complete the survey, place it in the envelope, seal the envelope, and return it to school with their son or daughter. Once at school, sealed envelopes were placed in a central storage container until they were retrieved by the research team.

The survey asked the student's first name, last name, date of birth, gender, grade level, and ethnicity. These data, along with the school the student was enrolled in were used to link survey data to achievement records. Questions were derived from prior reviews of educational research identifying family variables that predict student achievement. The survey questions and response options are presented in Table 21.

Table 21: *Family Demographic Survey Items*

Item	Response Options
Has the child always lived with the same parent (mom OR dad) since they were born?	Yes No
Has the child always lived with the same 2 parents (mom AND dad) since they were born?	Yes No
Marital status	Married Separated Divorced Never married Widowed
Number of adults that live in the home	Grid Number
Number of children that live in the home	Grid Number
Age of mother when child was born	Grid Number
Education level: mother	8th grade or less Some high school High school diploma Some college (at least 1 year) Vocational technical training College graduate Graduate professional degree
Education level: father	Same as mother
Annual family income	0-4,999 5,000-9,999 10,000-19,999 20,000-39,999 40,000 & up[1]
Number of times your family has moved since your child has been in school	0 1 2 3 or more
Number of times your child has changed schools since kindergarten	0 1 2 3 or more
How many activities outside of the home is your child involved in?	0 1 2 3 or more
How many of your child's friends' parents do you know?	0 1 2 3 or more
Do you have a working computer in your home?	Yes No
How many minutes per week do you spend with your child helping or talking about school work?	0-15 min 15-30 min 30 min-1 hour 1 hour +
How many times in a year do you visit your child's school for an event (do NOT include teacher conferences).	0 1 2 3 or more

Table note. 1. Due to a communication error the maximum value for family income was lower than would be desirable.

Results. The central research question for this data collection was the extent to which additional information from families would contribute variance in predicting student achievement that was unique from that contained in the educational databases available. As an initial step the variables that were included in the HLM models for each content area were entered as a block. The next stage of the analysis examined the ability of the family demographic survey to improve prediction of student achievement. Variables coded as yes or no were coded 1 for yes and 0 for no. Continuous variables (e.g., mother's age at student's birth) were entered as they were coded. Ordinal variables were coded from 1 to the top of their respective scale. Marital status was dichotomized into married and not married. The number of adults and children in the home was converted to the ratio of adults to children living in the home. Table 22 below provides the multiple *r*, the shared variance, and the number of participants for the educational variables model and the educational plus family variables models.

Table 22: Additional Contribution of Family Variables to Prediction of Achievement

Content	Educational Variables	Educational & Family Variables	*n*
Mathematics	.845 (.713)	.846 (.716)	2221
Reading	.804 (.646)	.806 (.650)	2220
Language Arts	.788 (.621)	.793 (.628)	2220
Science	.815 (.664)	.818 (.670)	2150
Social Studies	.786 (.618)	.790 (.623)	2150

Table note. The top value in each cell is the multiple correlation. The bottom value in parentheses is the squared multiple correlation.

Analyses revealed that the educational databases yielded predictions that were slightly more precise in the sample of responders than they were in the statewide analysis. The multiple *r* in the sample of responders was .001 to .031 points higher (mean = .012). This small difference suggests a high level of generalization from the statewide results to the survey sample in the overall performance of the predictors. The addition of all variables in the family survey accounted for such a small increment in additional variance (r^2 increased .003 to .007) that it is unlikely to have any meaningful impact on the estimation of TPP coefficients.

One of the questions that arose regarding the 2007 analyses was the ordering of variables among educational and family variables and the smallest efficient set for predicting achievement. Addressing that question (variable selection) would require a stepwise method. However, applying a stepwise method to such a large set of highly correlated and conceptually related variables is unlikely to produce replicable meaningful results (Thompson, 1995). The possible exception to this is the strong predictive power of prior year achievement in predicting current year achievement. Additional exploration of the family data is planned for a separate report.

The critical consideration for this report is the demonstration that the educational variables that are readily available account for a large proportion of current year achievement and that the addition of numerous family variables did not account for a meaningful increment in variance.

IX. Teacher ACT Scores at College Admission and Effectiveness

ACT test scores were obtained from the Louisiana Board of Regents and matched to the teacher data files in order to examine the impact of teacher candidates' educational attainment prior to entering teacher preparation on instructional effectiveness. These data could potentially address the issue of selectivity of admissions and the degree to which differences in TPP coefficients were the result of differential admissions. An initial examination of the seven alternative TPPs suggest that selection as indexed by the ACT scores is unlikely to have much explanatory power. The mean ACT scores for the programs for mathematics teachers ranged from 20.03 to 21.71, a very small proportion of the range of ACT scores. Similarly, the range of program means for language arts teachers ranged from 19.23 to 21.16. In neither case was any ordering of ACT mean score and TPP coefficients evident. For example, in mathematics, the two TPPs with the highest mean ACT had the lowest coefficient for their TPP. This is almost assuredly a chance fluctuation given how close all of the program mean ACT composite scores are. It is also interesting to note that these scores are consistent with the ACT scores of the bulk of the undergraduate programs prior to redesign and would fall in the center of that distribution.

Although the data suggest that mean ACT scores across do not differentiate these alternative certification programs, it remains possible that ACT scores would have some explanatory power at the level of individual teachers. HLM to assess teacher preparation programs based on the same specifications as those were implemented to examine this issue. The models excluded codes for the TPPs, but included all of the other variables described above. Models were examined separately for all teachers for whom an ACT score was available and separately for new teachers only. Separate analyses were conducted for each content area. At Level 2, teachers, the Composite ACT score was entered. In mathematics additional analyses were conducted using the Mathematics ACT score. Similarly, analyses were conducted in reading and language arts using the English ACT score.

Across all of the analyses, with the exception of one, the coefficient for ACT score was very small and not significant. The one exception was in mathematics where the coefficient was 0.23 ($p = .004$, grand mean centered). This coefficient suggests that a teacher whose Math ACT was 4 points above average would be predicted to have students whose test performance was 1 point above the projections based on student, class, and school data.

The authors want to emphasize that we do not interpret these results as suggesting that selection on variables such as ACT is unimportant in recruiting teacher candidates. Rather the data suggest that these alternative teacher preparation programs in Louisiana are so consistent in recruiting teacher candidates whose ACT scores are very near 20 or 21 that there is simply not enough variation on this issue to account for variance in TPP effectiveness. The data do suggest that at least in the domain of mathematics, despite the limited variability and despite the fact that teachers will have typically taken the ACT many years prior to entering teaching; that the mathematics knowledge as indexed by the ACT score was still a modest predictor of teacher effectiveness.

X. Teacher Certification

The research team also examined the impact of teacher preparation as indicated by teacher certification on teacher effectiveness. For purposes of this analysis all teachers who were uncertified, teaching on a temporary authority, or were teaching outside their area of certification were pooled. Subsequent analyses examining only teachers who were teaching under a temporary authority are planned, but in Louisiana that is a relatively small population. The effect estimates for teachers who were not certified in the area in which they were teaching is provided below in Table 23. All of the coefficients in the table below were statistically significant at $p < 0.001$ and demonstrate that teachers who are certified in the content area they are teaching are more effective than those who are not certified to teach that content.

Table 23: Impact of Teachers who are not Content Certified

Content	**Coefficient (CI)**
Mathematics	-3.50 (-4.70, -2.28)
Reading	-1.27 (-1.72, -0.82)
Language Arts	-4.09 (-4.70, -2.28)
Science	-1.58 (-2.34, -0.82)
Social Studies	-3.32 (-4.61, -2.03)

Table note. The top value in each cell is the coefficient for that content area. The bottom value in the bottom of the cell is the 95% confidence interval based on the SEM.

XI. Summary

Analyses were conducted to replicate and extend the prior statewide analyses for the 2004-2005 and 2005-2006 school years. Construction of the longitudinal database suggested that a sufficient quantity and quality of data appear to be available to support longitudinal analysis of educational inputs such as teacher preparation. For example, the 90% linkage rate for student data across years was very encouraging. The proportion of usable records was further decreased by issues such as student mobility and retention, but was above 72% of test takers in all content areas. It is important to acknowledge that as a result of screening measures used with the data, that these assessments are for teachers who remain in one school for the year, teaching the group of students who were promoted the prior year and who remain in that school the entire year. Although this approach selectively excludes teachers and students, it does permit comparison of TPPs in a common database. Examination of OLS regression analyses for the students who were eligible to contribute to the HLM VAA suggested sufficiently strong prediction of current year achievement to support the analysis.

The following points are primary findings of each stage of the analyses.

1. The ordinary least squares regressions demonstrate a strong relationship between prior year achievement and current year achievement in the content area. Adding achievement in the four other domains strengthened that relationship as did adding student level demographics and attendance data.

2. The mixed linear models developed for each of the content areas shared a great deal in common. Prior achievement, special education disability status, Section 504 entitlement, receipt of free/reduced price lunch, giftedness, gender, and student absences consistently entered the equations. Being African American was the only ethnicity code that consistently entered models and it loaded negatively in four of five content areas.

3. VAA of TPP was conducted across 2004-2005, 2005-2006, and 2006-2007 academic years. Effect estimates were identified at all five performance bands. Some consistency in TPP effects was evident with three programs clustering in the top of the distribution and two clustering in the bottom across content areas. For the two programs clustering at the bottom of the distribution among the programs included in this report, it is important to bear in mind that their program estimates were in the average range for new teachers in four of five content areas. Results were generally consistent at the level of performance bands across the 2007 and 2008 reports, with one TPP exhibiting a decline of one level across all content areas.

4. Consistent with the 2007 report, the correlation between the PSM derived estimates and the whole state derived estimates (.98 to .99) were high enough as to suggest there is little practical advantage to the PSM procedure for these data.

5. Examination of family demographic data in a sample of more than 2000 students found that the demographic variables increased variance shared with student achievement test scores to a very small degree that was unlikely to substantively affect the VAA of TPP. This is not to argue that family factors are unimportant; clearly, they are. They may simply share so much variance with the data already in the educational databases that they would add little to the assessment.

6. Examination of the ACT data suggested several initial findings. The TPPs included in this report were so consistent in the ACT scores of their graduates that selection as indexed by ACT scores is unlikely to have much explanatory value. Generally, ACT scores for new teachers in Louisiana clustered near 20 or 21. The only area in which ACT scores were predictive of subsequent teacher effectiveness was mathematics. It is noteworthy however that an assessment of mathematics content knowledge completed prior to admission to teacher preparation would predict effectiveness teaching mathematics years later.

7. Examination of the impact of teacher preparation as indexed by certification found that teachers who were not content certified were less effective than content area certified teachers. This difference was particularly large for language arts, mathematics, and social studies.

In summary, the data suggest that differences in TPP effectiveness are detectable using data pooled across multiple school years.

During the current academic year, data describing the TPPs will be made available to the VAA-TPP team that will permit examination of the degree to which program characteristics are associated with their impact on student attainment. Additionally, the 2009 report should be the first occasion in which post redesign data for undergraduate programs will become available in sufficient quantities to permit inclusion of those programs. This should be very interesting both in terms of permitting examination of the largest providers of new teachers and for examining the effects of the redesign of teacher preparation.

References

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29,* 37-65.

Coleman, J. S. (1989). The family, the community, and the future of education. In W. J. Weston (Ed.), *Education and the American family: A research synthesis.* (pp. 169-185). New York University Press: New York.

Chatterji, M. (2006). Reading achievement gaps, correlates, and moderators of early reading achievement: Evidence from the Early Childhood Longitudinal Study (ECLS) kindergarten to first grade sample. *Journal of Educational Psychology*, *98*, 489-507.

Downer, J. T., & Pianta, R. C. (2006). Academic and cognitive functioning in first grade: Associations with earlier home and child care predictors and with concurrent home and classroom experiences. *School Psychology Review, 35,* 11-30.

Goldhaber, D. & Brewer, D. (1997). Why don't schools and teachers seem to matter? Measuring the impact of unobservables on educational productivity. *Journal of Human Resources, 32,* 505-523.

Hill, H. C., Rowan, B. & Ball, Lowenberg, D. L. (2005). The effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal, 42,* 371-406.

Hill, N. E., & Craft, S. A. (2003). Parent-school involvement and school performance: Mediated pathways among socioeconomically comparable African American and Euro-American families. *Journal of Educational Psychology, 95,* 74-83.

Hong, G., & Raudenbush, S. W. (2008). Casual inference for time-varying instructional treatments. *Journal of Educational and Behavioral Statistics, 33,* 333-362.

Lockwood, J. R., & McCaffrey, D. F. (2007). Controlling for individual level heterogeneity in longitudinal models, with applications to student achievement. *Electronic Journal of Statistics, 1,* 223-252.

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value added accountability. *Journal of Educational and Behavioral Statistics, 31,* 35-62.

Martineau, J. A., Subedi, D. R., Ward, K. H., Li, T., Diao, Q., Drake, S., Kao, S. Li, X., Lu, Y., Pang, F., Song, T., & Zheng, Y. (2007). Non-linear unidimensional scale trajectories through multidimensional content spaces: A critical examination of the common psychometric claims of unidimensionality, linearity, and interval-level measurement. in R. W. Lissitz (Ed.) *Assessing and modeling cognitive development in school: Intellectual growth and standard setting.* JAM Press, Maple Grove, MN.

McCaffrey, D. F., Lockwood, J. R., Kortez, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability.* Santa Monica, CA: RAND corporation.

McCaffrey, D. F., Lockwood, J. R., Kortez, D. M., Louis, T. A., & Hamilton, L. S. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29,* 67-102.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models.* New York: Wiley.

Noell, G. H. (2004). *Technical report of: Assessing teacher preparation program effectiveness: A pilot examination of value added approaches.* Retrieved from http://asa.regents.state.la.us/TE/technical_report.pdf

Noell, G. H. (2005). *Assessing teacher preparation program effectiveness: A pilot examination of value added approaches II.* Retrieved from http://asa.regents.state.la.us/TE/value_added_model.

Noell, G. H. (2006). *Annual report of: Value added assessment of teacher preparation.* Retrieved from http://asa.regents.state.la.us/TE/value_added_model.

Noell, G. H. & Burns, J. L. (2006). Value added assessment of teacher preparation: An illustration of emerging technology. *Journal of Teacher Education, 57,* 37-50.

Noell, G. H., Porter, B. A., & Patt, R. M. (2007). Value added assessment of teacher preparation in Louisiana: 2004-2007. http://www.regents.state.la.us/Academic/TE/Value%20Added.htm.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*, 237-257.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd Ed.). London: Sage.

Reckase, M. D. (2004). The real world is more complicated than we would like. *Journal of Educational and Behavioral Statistics, 29,* 117-120.

Roderick, M., Jaccob, B. A., & Bryk, A. S. (2002). The impact of high-stakes testing in Chicago on student achievement in promotional gate grades. *Educational Evaluation & Policy Analysis, 24,* 333-357.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the *Prospects* study of elementary schools. *Teachers College Record, 104,* 1525-1567.

Rubin, D. B., & Thomas, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association, 95,* 573-585.

Sanders, W. L., & Horn, S P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12,* 247-256.

Seltzer, M. H., Frank, K. A., & Bryk, A. S. (1994). The metric matters: The sensitivity of conclusions about growth in student achievement to choice of metric. *Educational Evaluation and Policy Analysis, 16,* 41-49.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal, 113,* 3-33.

Tekwe, C. D., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., Ariet, M., Fisher, T., & Resnick, M. B. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics, 29,* 11-37.

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here: A guidelines editorial. *Educational and Psychological Measurement, 55,* 525-534.

Wayne, A. J., & Youngs, P. (2003). Teacher characteristics and student achievement gains. *Review of Educational Research, 73,* 89-122.

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin, 91,* 461-481.