

Annual Report of:

Value Added Assessment of Teacher Preparation

George H. Noell, Ph.D.
Department of Psychology
Louisiana State University

August 23, 2006

Acknowledgements

This report is based upon data provided by the Division of Planning, Analysis, and Information Resources. The author would like to thank Allen Schulenberg, Robert Kaufman, Kelvin LaCroix, Steve Gunning, Sam Pernici, Roth Aymond, and David Elder (Division Director) without whom this work would not have been possible.

Any errors, omissions, or misstatements contained herein are entirely the responsibility of the author. Any conclusions proffered are the responsibility of the author 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-05/06-VAA-02 from the Louisiana Board of Regents.

Table of Contents

Cover Page	1
Contents	2
Abstract	3
I. Overview	4
II. Data Merging Process	6
III. Preliminary Analyses	6
IV. Linking Students and Teachers.....	12
V. Building the Base Model of Student Achievement Prior to VAA.....	12
VI. Assignment of Teachers to Groups	21
VII. VAA of Teacher Preparation.....	23
VII. Effects of Different Normative Comparison Bases.....	31
VIII. Reliability of VAA Estimates	31
IX. Teacher ACT Scores at College Admission and Effectiveness.....	33
X. Secondary Analysis Findings.....	33
IX. Summary.....	34
References.....	38

Abstract

Value Added Assessment of Teacher Preparation

Analyses were conducted examining the feasibility of using Louisiana's student achievement, teacher, and curriculum databases to assess the efficacy of teacher preparation programs within Louisiana. 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 mixed linear models were developed to predict student achievement based upon prior achievement, student demographic factors, and classroom level covariates. The model nested students within teachers and teachers within schools. The model included effects for teachers and schools. Separate models were developed for each content area. These models were used to assess the efficacy of teacher preparation programs. The initial Value Added Assessment (VAA) of teacher preparation programs assigned programs to one of three categories. Level III was used to describe programs whose graduates were similar to experienced teachers in their effectiveness. Level II was used to describe universities for whom the data suggested a negative effect, but for whom the small number of new graduates or the variability of scores created more uncertainty. Level I was used to describe universities whose graduates were statistically significantly less effective than experienced teachers. Universities were identified that were at all three levels in all content areas. One issue of particular note for future research was the general imprecision of estimates of teacher program effects as suggested by the relatively large standard errors of measurement. This finding suggests that comparisons among new teacher preparation programs are problematic in the context of the obtained results. However, the availability of additional data in future years and examination of additional analytic options may provide means of improving upon these initial results. A number of issues remain for future research. These include issues of the graduates of particular preparation programs clustering in particular school districts and the use of models that included latent variable designs. However, these data suggest that even after including students' prior achievement, students' demographic variables, classroom context variables, and school building level effects, it is still possible to detect differences in teacher effectiveness for new teachers that differ across some university preparation programs. Much additional work remains to resolve technical and methodological issues to develop a maximally useful assessment approach.

Assessing Teacher Preparation Program Effectiveness: A Pilot Examination of Value Added Approaches

I. Overview

One of the enduring observations of educators and assumptions of policy makers has been that teachers influence the educational attainment of children. This long held belief has been supported by data from a considerable variety of divergent studies. To state the obvious: Teachers and teaching matter. This observation has in turn led to research literature regarding means of improving teaching, means of assessing teaching, and means of strengthening teachers' knowledge and skills. At the larger systemic level an academic skills as outcomes approach and high stakes standardized test-based accountability appear to have achieved ascendancy (Roderick, Jacob, & Bryk, 2002; Sheldon & Biddle, 1998). Although considerable forces have converged to create the current emphasis on measurable skills as outcomes when assessing educational units, it is important to acknowledge that support for this shift in policy is far from unanimous.

Despite an expanding literature examining methods of using standardized test-based accountability data to evaluate educational programs, broadly accepted approaches to using the data for this purpose have not yet emerged (Rowan, Correnti, & Miller, 2002; Thompson, 2004). One of the essential features of the challenge is that an end points as outcomes only model appears to be insufficient (McCaffrey, Lockwood, Kortez, & Hamilton, 2003). A highly effective teacher or curriculum can be provided to students and these students may still not reach the end point that is desired simply because they started out so far behind. The logical approach that has been the focus of increasing study has been to obtain a pretest and assess change (Ballou, Sanders, & Wright, 2004; McCaffrey, Lockwood, Kortez, Louis, & Hamilton, 2004; Sanders & Horn, 1998). Although considerable progress has been made in developing empirical methodologies in this area, some issues remain unresolved. For example, the distinction between the hypothetical versus the actual benefits of employing some complex methodologies versus somewhat simpler ones remains an area of active research (Tekwe et al., 2004). Similarly considerable disagreement exists regarding the most appropriate and effective means of dealing with factors such as race and poverty (Ballou et al., 2004; Webster & Mendro, 1997). Because of the considerable variance shared between demographic variables and prior achievement it is not clear that they substantially improve predictive accuracy, if prior achievement data are already accounted for (Ballou et al.).

Interestingly, although there is general unanimity that teachers matter, there has been little emphasis on the assessment of the source of teachers as it is related to student achievement. This is particularly interesting, because if teachers matter and the workforce of teachers can be strengthened during the lengthy preparation process (typically 2-4 years) then the impact on education would be expected to be considerable. However, new research in Louisiana has begun to examine how the efficacy of teacher preparation might be assessed through its ultimate outcome: the achievement of students taught by new completers of teacher preparation programs (Noell, 2004; Noell, 2005, Noell & Burns, 2006). These initial studies have suggested that given sufficient data, it may be possible to detect the impact of teacher preparation programs on students' achievement when those students are taught by recent program completers.

Assessing the effectiveness of newly prepared teachers is a critical challenge confronting universities, school districts, and states (Roderick, Jaccob, & Bryk, 2002; Thompson, 2004). The long standing societal desire to provide all children well prepared teachers, as well as the newly emerging emphasis on measurably effective schools and teachers evident in the No Child Left Behind Act of 2001 point toward the need for states to assess the effectiveness of teacher preparation programs. However, the relatively large number of new teachers, their geographic dispersion following graduation, the challenges associated with large-scale collection of valid measures, and the finite resources available have placed limits on what approaches have been practical for universities to pursue in assessing new teacher effectiveness. The most obvious indicator, the learning of K-12 students who are taught by new teachers, is challenging at both a pragmatic and conceptual level. At a pragmatic level, collecting student achievement data in thousands of classrooms distributed across a state is an enormous and expensive undertaking. Additionally, even if those data were readily available, developing an analytic model that permits meaningful comparisons among groups of new teachers based upon student achievement is an extremely challenging task conceptually (Ballou et al., 2004; McCaffrey et al., 2004; Tekwe et al., 2004). Interested readers might choose to consult Noell and Burns (2006) for a more detailed discussion of some of the practical, methodological, and policy issues surrounding assessing teacher preparation based upon student achievement data derived from standardized test performance. Additionally, the excellent special issue of the *Journal of Educational and Behavioral Statistics* (Spring, 2004) provides a detailed examination of the methodological and epistemological issues relevant to value added assessment of teachers and schools. It is worth noting that some of the issues relevant to the assessment of teacher preparation are distinct from the issues relevant to the assessment of teachers and schools.

Prior Pilot Work

Two years of pilot work were completed prior to the current year's research. The pilot research was completed using the routine student achievement assessment databases maintained by the State of Louisiana, teacher databases that were available, and a new data system linking teachers, students and courses. This new data system was pilot tested in 10 school districts for each of the two pilot years. Eight of those pilot school districts were the same for both pilot years. For the pilot data, the year to year association between the educational assessments used in Louisiana was sufficiently strong that the creation of a longitudinal analysis model appeared to be appropriate. In the first year, three analytic models were examined and all suggested similar conclusions (Noell, 2004). Based upon review of the data, the initial recommendation was made to pursue a hierarchical linear model or linear mixed model based upon their flexibility, their match to the natural structure of the data, and their potential for providing powerful analytic tools. Both prior years' pilot analyses found some instances in which new graduates from specific university preparation programs were similar to experienced teachers and some instances in which new graduates of specific teacher preparation programs were statistically significantly less effective than experienced teachers (Noell, 2004; Noell, 2005). The follow-up analyses in pilot year two replicated the first year findings with regard to most substantive issues relevant to the assessment of teacher preparation.

This technical report describes key findings of the first year of a multiyear research effort to examine the assessment of teacher preparation through value added assessment (VAA). The initial model development work based upon all 68 school districts in Louisiana is described herein. Additionally, preliminary analyses were performed using the model developed for the State with the data for the pilot school districts as well as the current year data to examine the stability of an estimate of teacher effectiveness across years and the influence of class size on this estimate.

II. Data Merging Process

The target year of teaching assessed was the 2004-2005 academic year. Data contributing to this analysis were drawn from the curriculum database linking students and teachers for 2004-2005 and spring standardized testing assessments (*ITBS* and *LEAP-21*) for spring 2004 and 2005. Additional data drawn from student databases and teacher certification databases were merged with the database. Initial work was undertaken to resolve duplicate records and multiple partially complete records that described the same student. Following this work, the *ITBS* and *LEAP-21* 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. Those records that did not uniquely match at any stage were retained as isolated records of student performance. The details of this process are available from the author.

At the end of this process, the data set contained 517,533 records, each representing 1 student in grades 3 through 11. Z-scores were then calculated based on the *LEAP-21* and *ITBS* scaled scores for English Language Arts (ELA), Mathematics, Science, and Social Studies within each grade level and year.

The target population for the analyses reported herein is students who were in grades 4 through 9 in 2005. Of the available students for 2005, 93.8% of student records were linked across years. Given the realities of students' moving out of state, moving into and out of private schools, spoiled tests, and clerical errors, this is a very encouraging level of matching. The potential pool of students for analysis was further limited to students who were promoted at the end of the 2004 school year. The analyses focused on students who were promoted because the meaning of the year-to-year comparisons of test performance is different for students who took the same test and repeated their grade than those who were promoted. This resulted in a potential target pool of 286,223 students for analysis. A further portion of these students were lost because two school districts did not provide the State Department of Education with course enrollment data. This is anticipated to be a one year anomaly resulting from this being the first year of the statewide collection of teacher-student-course linkage data.

III. Preliminary Analyses

Prior to pursuing examination of approaches to implementing a VAA of teacher preparation programs with Louisiana's achievement data, a series of statewide ordinary least squares (OLS) regression analyses were conducted to examine general patterns in

the data. Progressively more variables were employed as predictors and the multiple correlation between achievement in 2003 and predictor variables was examined. Initially, students who were in grades 4 through 9 in the spring of 2005, who took either the *ITBS* or *LEAP-21*, and were promoted at the end of the 2004 school year were identified as eligible for inclusion.

Table 1: *English-Language Arts Statewide Regression Analyses for 2005*

Predictors	Multiple correlation	Number of Students
Z-score: ELA 2004	.767	283,767
Z-scores 2004 achievement	.792	282,785
Z-scores 2004 achievement Student demographic factors	.803	282,785
Z-scores 2004 achievement School demographic factors	.793	279,537
Z-score: ELA 2003 & 2004	.815	207,363
Z-scores 2003 & 2004 achievement	.825	206,727
Z-scores 2003 & 2004 achievement Student demographic factors	.830	206,727
Z-scores 2003 & 2004 achievement School demographic factors	.825	204,611
Z-score: ELA 2002 – 2004	.832	152,322
Z-scores 2002 – 2004 achievement	.840	151,944
Z-scores 2002 – 2004 achievement Student demographic factors	.844	151,944
Z-scores 2002 – 2004 achievement School demographic factors	.840	150,486

Note. *Year achievement* includes the Z-scores for ELA, mathematics, science, and social studies. *Student demographic factors* included were free lunch status, gifted status, other special education status, limited English proficiency status, gender, and minority status. *School demographic factors* included the number of students at the school, percentage of students receiving free/reduced cost lunch, percentage of students who were minorities, percentage of students who were male, percentage of students identified as disabled, percentage of students identified as gifted, and percentage of students identified as having limited English proficiency.

Table 2: *Mathematics Statewide Regression Analyses for 2005*

Predictors	Multiple correlation	Number of Students
Z-score: Math 2004	.789	283,627
Z-scores 2004 achievement	.806	282,728
Z-scores 2004 achievement Student demographic factors	.813	282,728
Z-scores 2004 achievement School demographic factors	.809	279,373
Z-score: Math 2003 & 2004	.830	207,288
Z-scores 2003 & 2004 achievement	.837	206,702
Z-scores 2003 & 2004 achievement Student demographic factors	.839	206,702
Z-scores 2003 & 2004 achievement School demographic factors	.838	204,506
Z-score: Math 2002 - 2004	.843	152,274
Z-scores 2002 – 2004 achievement	.848	151,933
Z-scores 2002 – 2004 achievement Student demographic factors	.849	151,933
Z-scores 2002 – 2004 achievement School demographic factors	.849	150,410

Note. *Year achievement* includes the Z-scores for ELA, mathematics, science, and social studies. *Student demographic factors* included were free lunch status, gifted status, other special education status, limited English proficiency status, gender, and minority status. *School demographic factors* included the number of students at the school, percentage of students receiving free/reduced cost lunch, percentage of students who were minorities, percentage of students who were male, percentage of students identified as disabled, percentage of students identified as gifted, and percentage of students identified as having limited English proficiency.

Table 3: *Science Statewide Regression Analyses for 2005*

Predictors	Multiple correlation	Number of Students
Z-score: Science 2004	.726	282,089
Z-scores 2004 achievement	.772	281,986
Z-scores 2004 achievement Student demographic factors	.781	281,986
Z-scores 2004 achievement School demographic factors	.778	279,024
Z-score: Science 2003 & 2004	.779	206,391
Z-scores 2003 & 2004 achievement	.803	206,292
Z-scores 2003 & 2004 achievement Student demographic factors	.807	206,292
Z-scores 2003 & 2004 achievement School demographic factors	.805	204,301
Z-score: Science 2002 - 2004	.800	151,723
Z-scores 2002 – 2004 achievement	.819	151,618
Z-scores 2002 – 2004 achievement Student demographic factors	.821	151,618
Z-scores 2002 – 2004 achievement School demographic factors	.821	150,257

Note. *Year achievement* includes the Z-scores for ELA, mathematics, science, and social studies. *Student demographic factors* included were free lunch status, gifted status, other special education status, limited English proficiency status, gender, and minority status. *School demographic factors* included the number of students at the school, percentage of students receiving free/reduced cost lunch, percentage of students who were minorities, percentage of students who were male, percentage of students identified as disabled, percentage of students identified as gifted, and percentage of students identified as having limited English proficiency.

Table 4: *Social Studies Statewide Regression Analyses for 2005*

Predictors	Multiple correlation	Number of Students
Z-score: Social Studies 2004	.691	281,954
Z-scores 2004 achievement	.749	281,885
Z-scores 2004 achievement Student demographic factors	.755	281,885
Z-scores 2004 achievement School demographic factors	.753	278,925
Z-score: Social Studies 2003 & 2004	.742	206,318
Z-scores 2003 & 2004 achievement	.776	206,234
Z-scores 2003 & 2004 achievement Student demographic factors	.778	206,234
Z-scores 2003 & 2004 achievement School demographic factors	.777	204,248
Z-score: Social Studies 2002 - 2004	.765	151,650
Z-scores 2002 – 2004 achievement	.791	151,572
Z-scores 2002 – 2004 achievement Student demographic factors	.791	151,572
Z-scores 2002 – 2004 achievement School demographic factors	.791	150,218

Note. *Year achievement* includes the Z-scores for ELA, mathematics, science, and social studies. *Student demographic factors* included were free lunch status, gifted status, other special education status, limited English proficiency status, gender, and minority status. *School demographic factors* included the number of students at the school, percentage of students receiving free/reduced cost lunch, percentage of students who were minorities, percentage of students who were male, percentage of students identified as disabled, percentage of students identified as gifted, and percentage of students identified as having limited English proficiency.

The most striking outcome of the preliminary statewide regression analyses was the strong relationship between achievement scores across years. The importance of

demographic factors attenuates as more years of achievement data become available. It is clear that in this preliminary analysis, school building level demographic variables contributed less variance than student level demographic factors. Even at one year of prior achievement data combined with student level demographic factors, a very strong multiple correlation was obtained. Across all content areas, the mean multiple correlation for prior year achievement plus demographics was .79 for English Language Arts (ELA) and for mathematics the mean was .81. These data again support the potential of Louisiana's educational databases to support longitudinal data analyses such as a VAA of teacher preparation. They also suggest that a single year's prior test scores may be sufficient when that prior year's data includes four content areas and can be augmented by demographic variables that are commonly available in education.

IV. Linking Students and Teachers

Following preliminary linking of data and analyses, the student achievement data were linked with the data connecting students to courses and courses to teachers. In addition, selected data from the Profile of Educational Personnel (PEP) and the certification database provided by the Louisiana Department of Education's Division of Planning, Analysis, and Information Resources were linked to teachers and the longitudinal educational achievement database. These data permitted identification of new teachers. Additionally, in order to contribute to these analyses, each student had to be enrolled in the same school in the fall of 2004 and at the spring assessment of 2005.

Course codes were collapsed into groups that were associated with specific test areas (i.e., ELA, mathematics, science, and social studies). For example, English I was associated with ELA tests and Life Science with science tests. If the student did not have a specific teacher identified for a particular content area, but had a teacher identified by a broad range of content areas (e.g., the code elementary grades), then the teacher in the broad category was linked to that test outcome. Course codes that could not reasonably be linked to a standardized test (e.g., Jazz Ensemble) were dropped.

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

Following from the findings of the pilot investigations, the educational assessment data were analyzed using hierarchical linear models (HLM; McCulloch & Searle, 2001; Raudenbush & Bryk, 2002). HLM or mixed linear models have several important advantages over traditional analytic approaches. First, they readily capture the grouping of students within classrooms. Second, they permit appropriate modeling of variables at multiple levels such as student, teacher, and school. Third, they provide a model in which estimates of teacher effectiveness can be adjusted to account for unreliability of estimates.

The modeling employed in this year used a 3 layer structure. Students were grouped within teachers' classes who were, in turn, grouped within schools. This differs from the prior pilot work in that a school building level was included in the model. The decision to add a school layer to the model was based upon several factors. First, there was an interest in including the effect of higher level organizations (districts and/or schools) on students' educational outcome. Second, in a series of preliminary analyses

the proportion of variance that was distributed among students, teachers, schools, and school districts was examined. The critical findings of this analysis were twofold. First, the variance component within school districts was relatively trivial. This makes some conceptual sense given the reality that schools within school districts can be quite heterogeneous. Second, the variance component associated with schools while small, ranging from 2.6% for mathematics to 3.2% for ELA, is likely to be of some importance. Of the variability that was associated with schools and teachers, approximately 20% of that variance was associated with schools and 80% of that variance was associated with teachers based upon the standardized achievement data for the State of Louisiana for the 2004-2005 school year.

Table 5 presents the findings for ELA and mathematics for teachers, schools, and school districts in three different models for the nesting of teachers. These models included prior achievement, student demographics, and classroom demographics as predictors of student achievement. The data presented below were based upon the final models developed for each content area, which are described in detail later in this report.

Table 5: *Preliminary Shared Variance for Different Nesting Models that include Student and Classroom Level Predictors*

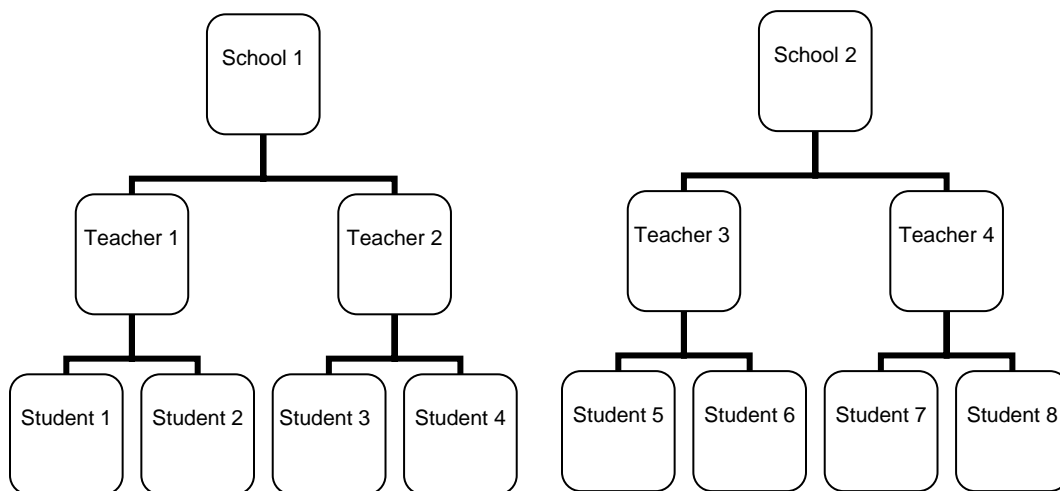
Content	Variance Component	Model		
		Students within Teachers	Students within Teachers within Schools	Students with Teachers within School Districts
Mathematics	Teachers	13.7%	10.9%	13.2%
	Schools		2.6%	
	School Districts			0.5%
	Students	86.3%	86.5%	86.3%
ELA	Teachers	14.5%	11.3%	13.9%
	Schools		3.2%	
	School Districts			0.5%
	Students	85.5%	85.5%	85.6%

It is interesting to note the relationship between the estimated teacher/school effects and the duration of students' educational careers. For the students contributing to this analysis, the school year studied is 1 year out of 7.5 years on average of total school attendance or 13% of their school career to date. This figure corresponds closely to the proportion of variance the data suggest is attributable to teachers/schools in this one year

(approximately 14%). It is important to note that although the average effect of one teacher in one year is not large, that over several years the cumulative effects of teachers would be expected to be substantial (e.g. Sanders & Horn, 1998).

It is also *critically important* to recognize that this estimate is an estimate of the current year's teacher's contribution to the all of the learning assessed in the current year's test. An assessment model that removes prior years' achievement and attempts to isolate only current year achievement would undoubtedly result in a much larger estimate of the contribution of teachers for both conceptual and pragmatic reasons (see McCaffrey et al., 2003 and Rowan et al., 2002). Figure 1 below depicts the final nesting structure that was employed.

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



Building the current models. The model development began with no prior assumptions beyond those that are common to mixed linear models. The modeling approach was somewhat parallel to Tekwe and colleagues (2004) in general strategy. The approach was replicated across ELA, mathematics, science, and social studies. 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. Next, the students' prior year's achievement in ELA, mathematics, science, and social studies were entered as a block as fixed effects. All effects were significant in all content areas and were retained. Next, nine demographic variables were entered as a block. The nine demographic variables are presented in Table 6 along with the percentages of the sample for which the demographic variable was coded as true. Variables were then removed one at a time in order of the lowest t value until all remaining effects were significant.

It is worth noting that addition of student demographic variables, while statistically significant in a context with in excess of 200,000 students for each analysis, accounted for very little variability in student achievement. For example, demographic

factors accounted for 4% of the variability in student achievement after accounting for prior achievement in ELA. In mathematics, adding demographic factors to prior achievement accounted for an additional 1.7% of the variance. In short, at the student level for Louisiana's data, the bulk of the variability in any one year is shared with the prior year's achievement in the four curricular content areas.

Table 6: *Student Level Demographic Variables*

Variable	Percentage of Sample
Receiving Free/Reduced Lunch	62.0%
Special Education	13.3%
Gifted	5.8%
Limited English Proficiency	0.9%
Gender (Male)	50.3%
Native American	0.7%
Hispanic	1.7%
Asian American	1.3%
African American	46.8%

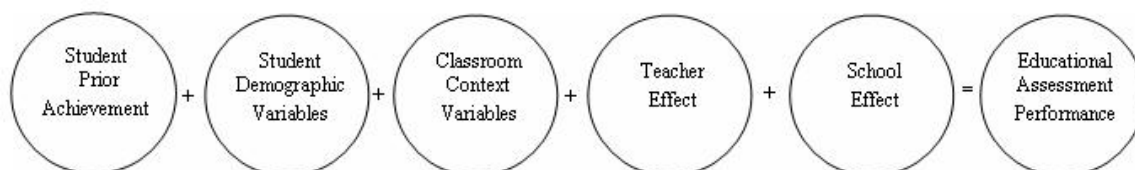
Once a model for student level achievement was developed, several classroom variables were examined. These variables were entered at the teacher level and were conceptualized as contextual factors that may moderate student achievement in addition to specific effects of the students' teachers. The variables that were examined are presented in Table 7.

Table 7: *Classroom Level Demographic Variables*

Variable
Percentage of students who were male
Percentage of students who were minorities
Percentage of students who received free or 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

As with the student level demographic factors these classroom context variables were entered as a block and then removed one at the time in order of smallest t value for the coefficient. Once all effects were significant the final model for that content area was finalized. Classroom level variables accounted for a modest portion of the variance that was associated with teachers. Classroom covariates accounted for 5% of teacher variance for ELA and 7.4% of the teacher variance for mathematics. It is important that these estimates are based on a model that already includes *extensive* data about the students individually. These data suggest that classroom composition accounts for a modest portion of the variability in teacher effects, once detailed information about the students individually is already included. These compositional effects would certainly be much larger if student level data were not already included.

Figure 1 below presents a graphic representation of the model of student achievement examined in this study.

Figure 2: *Factors Included in the Models of Student Achievement*

The following tables present the variables that were retained at the student and teacher levels for each content area prior to consideration of teacher preparation effects. In all cases models were developed for intercepts as outcomes with fixed effects at the

student level. The values presented in the tables are the final values that were obtained with teacher preparation programs included in the model. The coefficients for university preparation are presented in the section regarding the VAA of teacher preparation.

Table 8: *Hierarchical Linear Model for ELA Achievement*

Model Level	Variables Entered	Coefficient	(CI)
Student level variables	Gender (Male)	-9.2	(-9.5, -8.9)
	Hispanic American	1.2	(0.3, 2.2)
	Asian American	5.2	(4.1, 6.3)
	Free/reduced price lunch	-3.0	(-3.3, -2.7)
	Special Education	-10.8	(-11.4, -10.2)
	Gifted	8.9	(8.2, 9.6)
	Prior year ELA test	22.4	(22.1, 22.7)
	Prior year Math test	7.1	(6.8, 7.3)
	Prior year Science test	4.4	(4.1, 4.6)
	Prior year Social Studies test	5.1	(4.8, 5.3)
Classroom variables	% Boys	-0.5	(-0.9, -0.1)
	% Minority	-0.2	(-0.4, 0.0)
	% Free/reduced lunch	-0.5	(-0.8, -0.2)
	% Special Education	-0.5	(-0.8, -0.2)
	Mean prior achievement in ELA	-5.9	(-7.5, -4.3)
	Mean prior achievement in Math	6.3	(4.6, 8.0)

It is important to note that differences in how variables were scaled create the need for caution in comparing the coefficients across differing types of predictors. All effects describe the effect of the variable on predicted achievement in the units of the test: a mean of 300 with a standard deviation of 50 (the approximate scale of Louisiana's LEAP-21 test). 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 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 ELA achievement among the achievement predictors was his or her achievement in that domain the prior year. The coefficient for prior achievement in ELA was more than twice the value of any other variable's coefficient. Following that, giftedness, special education status, and gender all made similar contributions, with being male and in special education being negatively weighted. The remaining coefficients were slightly smaller but similar in magnitude, with the exception of being Hispanic American, which was quite small. It is important to recognize that the classroom level variables' contributions are entered in a context in which the individual level data are already present. All of the variables at the classroom level are in what would be the commonly expected direction other than mean prior

achievement in ELA. It may be that this variable provides a corrective measure when collections of extreme individual scores create an overly negative or overly positive prediction.

Table 9: *Hierarchical Linear Model for Mathematics Achievement*

Model Level	Variables Entered	Coefficient	(CI)
Student level variables	Gender	2.3	(2.0,2.6)
	African American	-5.1	(-5.5,-4.7)
	Native American	-1.4	(-2.8,-0.1)
	Asian American	6.4	(5.2,7.5)
	Free/reduced price lunch	-2.0	(-2.3,-1.7)
	Limited English Proficiency	4.6	(2.9,6.2)
	Special Education	-7.7	(-8.5,-7.0)
	Gifted	7.0	(6.3,7.8)
	Prior year ELA test result	6.5	(6.3,6.8)
	Prior year Math test result	26.3	(25.9,26.7)
	Prior year Science test result	3.9	(3.7,4.2)
Prior year Social Studies test result	2.1	(1.9,2.4)	
Classroom variables	Classroom % Boys	-0.8	(-1.2,-0.4)
	Classroom % Minority	-0.3	(-0.5,-0.1)
	Classroom % Free/reduced lunch	-0.5	(-0.7,-0.3)
	Classroom % Special Education	-0.5	(-0.8,-0.2)
	Classroom % Gifted	1.0	(0.8,1.2)

It is important to note that differences in how variables were scaled create the need for caution in comparing the coefficients across differing types of predictors. All effects describe the effect of the variable on predicted achievement in the units of the test: a mean of 300 with a standard deviation of 50 (the approximate scale of Louisiana's LEAP-21 test). 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 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.

Overall the results for mathematics are similar to those for ELA. The largest contributor in the prior achievement domain by far was the student's prior year achievement in mathematics. Beyond that, among the demographic variables, being gifted, being in special education, being Asian American, and being African American all contributed to similar degrees. Interestingly, for mathematics mean prior achievement for the students taught by that teacher did not remain in the final model. Several demographic variables all entered with similar relatively small weights.

Table 10: *Hierarchical Linear Model for Science Achievement*

Model Level	Variables Entered	Coefficient	(CI)
Student level variables	Gender	3.1	(2.7,3.4)
	African American	-8.6	(-9.0,-8.2)
	Free/reduced price lunch	-2.6	(-2.9,-2.2)
	Special Education	-4.8	(-5.5,-4.2)
	Gifted	7.9	(7.1,8.7)
	Number of Science Classes	1.0	(-0.2,2.2)
	Prior year ELA test result	5.5	(5.3,5.8)
	Prior year Math test result	8.2	(7.9,8.4)
	Prior year Science test result	15.0	(14.7,15.4)
	Prior year Social Studies test result	9.1	(8.9,9.4)
Classroom variables	Classroom % Boys	-0.5	(-0.8,-0.2)
	Classroom % Minority	-0.5	(-0.7,-0.3)
	Classroom % Free/reduced lunch	-1.0	(-1.3,-0.7)
	Classroom % Gifted	1.0	(0.7,1.3)
	Mean prior achievement in Math	3.5	(1.5,5.5)
	Mean prior achievement in Science	-2.8	(-5.0,-0.6)
	Mean prior achievement in Social Studies	-3.0	(-5.2,-0.8)

It is important to note that differences in how variables were scaled create the need for caution in comparing the coefficients across differing types of predictors. The results for science achievement roughly parallel the prior two content areas at the student level with prior achievement in science making the largest single contribution. However, the magnitude of the difference is somewhat smaller. Prior mathematics achievement, social studies achievement, being gifted, and being African American then formed a cluster of factors with similar weights. Several factors were significant at the classroom level, with the largest coefficients being associated with percentages of the class who were gifted, received free/reduced price lunch, and were minorities.

Table 11: *Hierarchical Linear Model for Social Studies Achievement*

Model Level	Variables Entered	Coefficient	(CI)
Student level variables	Gender	1.8	(1.4,2.1)
	African American	-3.0	(-3.4,-2.6)
	Free/reduced price lunch	-3.5	(-3.8,-3.2)
	Limited English Proficiency	2.5	(0.5,4.6)
	Special Education	-5.5	(-6.2,-4.8)
	Gifted	8.4	(7.6,9.1)
	Number of Social Studies Classes	2.3	(1.1,3.4)
	Prior year ELA test result	6.2	(6.0,6.5)
	Prior year Math test result	6.1	(5.8,6.4)
	Prior year Science test result	10.8	(10.5,11.1)
Prior year Social Studies test result	14.6	(14.2,14.9)	
Classroom variables	Classroom % Boys	-0.5	(-0.9,-0.1)
	Classroom % Minority	-0.5	(-0.7,-0.3)
	Classroom % Free/reduced lunch	-0.5	(-0.8,-0.2)
	Classroom % Gifted	0.5	(0.2,0.8)
	Mean prior achievement in Math	4.9	(2.8,7.0)
	Mean prior achievement in Social Studies	-4.2	(-6.5,-2.0)

It is important to note that differences in how variables were scaled create the need for caution in comparing the coefficients across differing types of predictors. The results for social studies achievement differed somewhat from the models above. While prior social studies achievement remained the largest single contributor, the coefficient for prior science achievement was relatively close in magnitude. As a group, the coefficients for prior ELA achievement, mathematics achievement, being identified as gifted, and being in special education were of a lower and similar magnitude. At the classroom level four demographic factors were retained as well as prior achievement in mathematics and social studies. All of the effects are in what would be the commonly expected direction, except for prior achievement in social studies for the students as an average. Similar to ELA, it may be that this variable provides a corrective measure when collections of extreme individual scores create an overly negative or overly positive prediction.

Summary. Generally the student level models were quite similar. For all content areas the largest single predictor was prior year's achievement in the target content and all other prior year achievement scores were statistically significant. Beyond that, gender, free/reduced lunch, special education, and gifted status were retained in the models for all four content areas. Identification as an African American student was retained in 3 of 4 models.

At the classroom level the models were more varied across content areas. This likely results from these variables accounting for considerably less of the variability in

student achievement and as a result being less stable. All of the classroom models included percentage of students who were boys, minorities, and received free/reduced lunch. The percentage of students who were gifted entered for 3 of 4 content areas and the percentage who were in special education entered for 2 of 4. Prior achievement in some content areas entered for 3 of 4 models, but the collection that entered and the direction of the loadings did not exhibit a clear pattern.

VI. Assignment of Teachers to Groups

The definition of a “new” teacher adopted in the prior pilot research was teachers in their first three years of teaching. This definition was adopted somewhat arbitrarily at the outset of the work guided by prior research, induction models, and common perceptions. However, it is not clear in an absolute sense that new teachers are those in their first three years of teaching. Secondary analyses were conducted to examine the mean teacher effect by years of teaching experience. These estimates were developed by applying the full model developed above to the data. From these estimates, the empirical Bayes residual for each teacher was extracted and used as an estimate of the degree to which that teacher’s effect differed from expectations based upon the available data. These estimates were then averaged by years teaching experience and plotted. The estimates for ELA and Mathematics are presented in Figures 3 and 4 below.

Figure 3

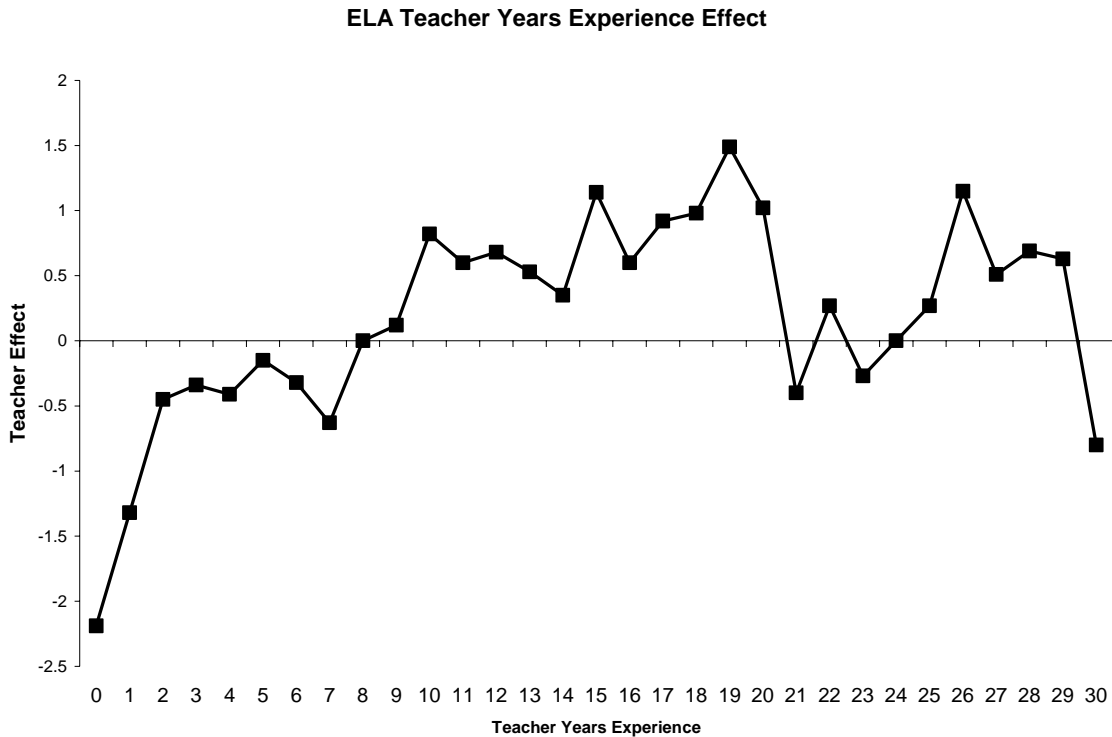
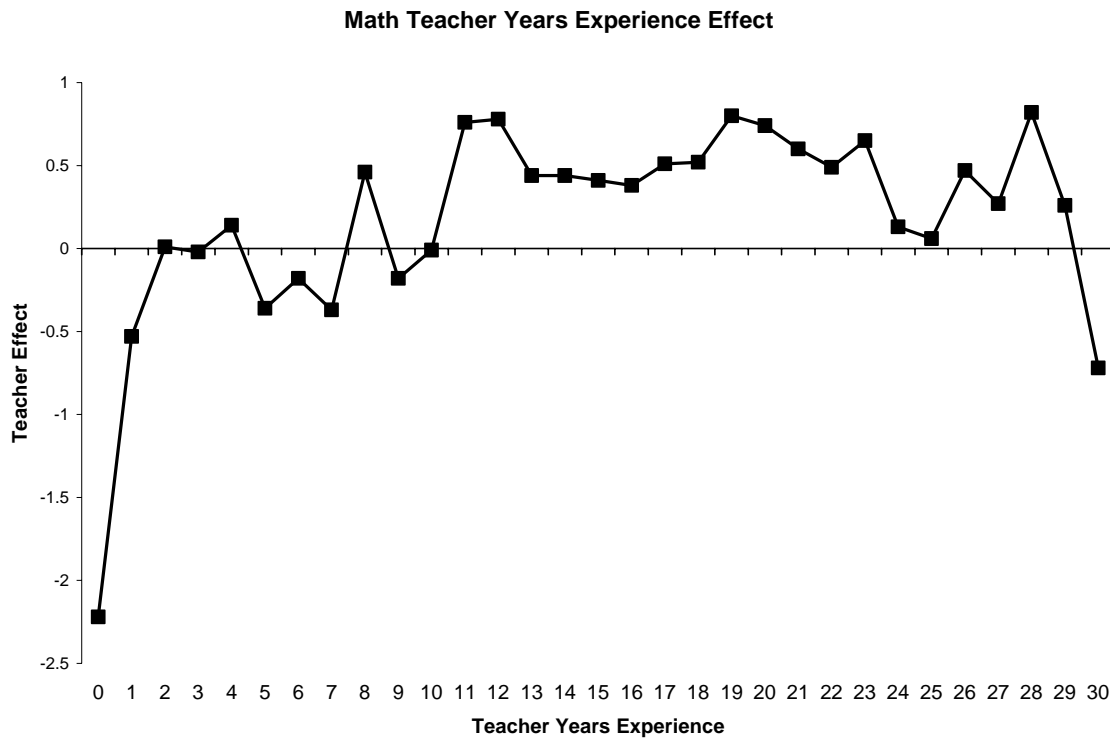


Figure 4



In ELA mean teacher effectiveness increased by approximately 1.5 points over the first three years of teaching. It then leveled off for several years with a second increasing stage of approximately 1.5 points occurring approximately between 10 and 20 years experience. The pattern for mathematics is quite similar with the initial gain being slightly more steep and the subsequent gain slightly more modest. Overall, while the gains are certainly not trivial, increases in teacher effectiveness with years of experience may be less dramatic than has been suggested by some policy makers and researchers.

It is important to acknowledge a relatively unique historical artifact in evaluating these data. Prior to 2003 Louisiana did not have requirements for ongoing professional development, either advanced degrees or job embedded, for the maintenance of professional licensure to teach. The enactment of this requirement is too recent to have had a substantial impact on these data. Based upon these findings the definition of new teacher adopted for this study was changed from the prior pilot work. A new teacher was defined as a first or second year teacher as described below.

Table 12: *Teacher Group Assignment*

Group	Criteria
New teachers	<ol style="list-style-type: none"> 1. First and second year teachers. 2. Holding a C or L1 certificate. 3. Received a university degree within 5 years of the start of school.
Regularly Certified Teachers	<ol style="list-style-type: none"> 1. Teachers with 2 or more years experience. 2. Holds an A, B, C, L1, L2, or L3 certificate.
Other	<ol style="list-style-type: none"> 1. Does not conform to any of the categories above.

All subsequent analyses were based upon this categorization combined with the teachers' degree granting institution.

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 the impact of teacher preparation programs. This step in the analysis, the VAA, examined the extent to which being a new program completer from a particular university preparation program was associated with improved or poorer student achievement. This was modeled at the teacher level by a series of dummy codes representing being a new program completer from a particular university. Alternative certification and regular undergraduate programs were modeled separately. In order for a program to be included in analyses it had to have at least 10 new graduates available for analysis. As a result of this rule no master's degree only programs were included in the analyses.

Implementing the planned statewide analysis and including the data for students who were enrolled in two teachers' classrooms resulted in a unique modeling challenge. Several software options were either investigated or tested. However, the scale of the analysis problem and the complexity of the analyses did not result in a satisfactory solution among the options initially tested. As a result a design option was selected. An analysis of the model was run twice. In each analysis students who had only one teacher were included each time. Eighty-three percent of students had only one teacher in ELA, 89% of students in mathematics, and 91% for both science and social studies had only one teacher. Students with two teachers were included in the first analysis linked to one teacher and in the second linked to the other teacher. The linking of teachers and students was divided across analyses such that teacher's links to students were equally represented across the two analyses (one half of the links per analysis). This design feature assigned all of the teacher effects to be assigned to the one teacher who taught students with a single teacher. It also divided the effect of teachers upon students between both teachers for students with two teachers. Reported effects are the means of the two analyses.

The impact of the programs was modeled on the scale of the current *LEAP-21* test due to its importance as the high stakes assessment for promotion in grades 4 and 8 as well as its disproportionate weight in School Performance Scores. The *LEAP-21* test for the target year 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. The number in brackets below the effect is the confidence interval for the effect.

University teacher preparation effects were classified as falling within four categories based upon the data. A Level IV program exhibited an effect on students taught by its graduates that *exceeded* the effect of experienced certified teacher at a level that was statistically significant. A Level IV program's new graduates were reliably more effective than experienced certified teachers. No Level IV programs were identified for this year for overall effects for grades 4-9. One program (University B, Undergraduate, Science) did exhibit Level IV effects within the grades 6-9 grade band. It is anticipated with additional years' data contributing to more precise estimates that it might be possible to identify a very few additional Level IV programs.

A Level III program was defined as a program whose estimated effect was close to that of experienced teachers. Their program effects were not statistically significantly different from experienced certified teachers and their estimated effect was within the 95% confidence interval of the mean experienced teacher effect based upon the mean standard error of measurement for teacher preparation programs. These programs' point estimates fell either slightly above to modestly below the effectiveness of new teachers. Generally, their point estimate was above the estimated effect for all first year teachers.

A Level II program was defined as a program whose effect was estimated to be negative and the negative effect too large to qualify as a Level III program. Additionally, the apparent effect was not statistically significant. These are programs whose data suggest a reason for concern, but whose number of graduates is so few or the variability of the performance of their graduates is so great that statistically significant effects were not obtained for this one year. It is anticipated that there would always be some programs in Level II, however, as more data from multiple year assessments of programs become available, it is likely that with greater precision some programs would move out of Level II to either Level I or III.

A Level I program was defined as a program whose effect on student achievement was negative and that effect was statistically significant.

The following preliminary descriptive labels are suggested for the levels: Level IV: Statistically Significantly More Effective than Experienced Teachers; Level III: Comparable to Experienced Teachers; Level II: Less Effective than Experienced Teachers; and Level I: Statistically Significantly Less Effective than Experienced Teachers.

Tables 13-16 below present the VAA estimates for ELA, mathematics, science, and social studies. The number of program completers per university program and content area are not being reported at this time because it became obvious to the research staff that many universities could readily be identified through their graduation patterns. Because the research is in a developmental stage and it is the intent of the research team

to maintain the anonymity of the programs until processes described herein are appropriately evaluated the number of graduates per program and content area are not reported at this time.

Table 13: *Teacher Preparation Program Effects: English Language Arts 2004-2005*

Level	University Teacher Preparation Program	Effect for Overall Achievement (CI)	Effect for Grades 4-6 (CI)	Effect for Grades 6-9 (CI)
III	Undergraduate Univ. B	-1.5 (-4.0, 0.9)	0.9 (-3.6, 5.4)	-3.3 (-5.4, -1.1)
III	Undergraduate Univ. F	-2.0 (-5.8, 1.7)	1.5 (-4.7, 7.7)	-5.6 (-10.7, -0.5)
III	Undergraduate Univ. H	1.9 (-2.8, 6.7)	3.3 (-5.6, 12.2)	0.1 (-2.5, 2.8)
III	Undergraduate Univ. I	-2.7 (-5.4, 0.1)	-3.5 (-8.2, 1.1)	-1.4 (-4.5, 1.6)
III	Undergraduate Univ. K	-1.5 (-8.7, 5.7)	-4.1 (-13.1, 4.8)	2.7 (-9.5, 14.9)
III	Undergraduate Univ. L	-2.4 (-6.5, 1.8)	-6.0 (-11.3, -0.6)	-0.3 (-5.2, 4.6)
III	Undergraduate Univ. M	-2.4 (-5.4, 0.7)	-3.7 (-8.2, 0.7)	-2.0 (-4.5, 0.5)
III	Alternative Cert. Univ. F	-2.6 (-7.1, 1.8)	ISD	ISD
III	Alternative Cert. Univ. I	-0.5 (-6.8, 5.8)	-17.7 (-33.1, -2.3)	3.6 (-0.7, 7.9)
III	Alternative Cert. Univ. M	-1.7 (-4.7, 1.4)	ISD	-2.4 (-5.1, 0.3)
III	Alternative Cert. Univ. O	1.8 (-6.2, 9.8)	ISD	ISD
II	Undergraduate Univ. A	-7.6 (-16.7, 1.6)	ISD	ISD
II	Undergraduate Univ. D	-3.9 (-8.1, 0.3)	ISD	ISD
II	Undergraduate Univ. E	-3.3 (-7.0, 0.4)	-4.4 (-9.9, 1.1)	-1.8 (-6.8, 3.2)
II	Undergraduate Univ. G	-2.9 (-12.4, 6.6)	ISD	ISD
II	Undergraduate Univ. J	-4.0 (-9.3, 1.4)	-4.4 (-11.7, 3.0)	ISD
II	Undergraduate Univ. N	-3.4 (-11.0, 4.2)	ISD	ISD

Table 13 (continued)

II	Alternative Cert. Univ. B	-3.0 (-7.0, 1.1)	ISD	-1.5 (-6.7, 3.7)
II	Alternative Cert. Univ. E	-2.9 (-11.7, 5.9)	-5.5 (-23.5, 12.5)	ISD
II	Alternative Cert. Univ. G	-7.6 (-16.5, 1.2)	ISD	ISD
I	Alternative Cert. Univ. L	-5.6 (-10.0, -1.2)	ISD	-2.7 (-8.2, 2.7)
I	Alternative Cert. Univ. P	-3.7 (-7.2, -0.2)	-4.0 (-10.5, 2.4)	-3.5 (-7.4, 0.4)
I	Alternative Cert. Univ. Q	-11.5 (-23.4, -0.3)	ISD	ISD

Note. The top number in each cell is the mean adjustment to student outcome that would be expected based upon a standard deviation of 50. The numbers in parentheses are the 95% confidence intervals. ISD indicates insufficient data in that cell to report results. The minimum number of new teachers per cell was set at 10.

Eleven programs were identified at Level III and these included both undergraduate programs and alternative certification programs. The mean effect for these programs ranged from -2.7 points on a students' test performance to +1.8. A relatively large number of programs fell in Level II, nine, whose effectiveness was uncertain. Many of these programs had relatively few program completers (less than 20) and the others exhibited considerable variability among their graduates. Three programs were identified as falling at Level I. They performed statistically significantly more poorly than experienced teachers. While the effects were negative and could make the difference between passing and failing one of the high stakes assessments for promotion/diploma conferral, they were generally not large negative effects. The effect for University Q may be a source of grater concern.

Comparison among teacher preparation programs is problematic due to the large 95% confidence intervals. For the all grades comparisons, the 95% confidence intervals for all programs overlapped. It is unclear at present whether using the narrower, less stringent 68% confidence intervals is advised. It is also possible that future research employing multiple year composites that will double the amount of data available would provide more precise and useful estimates. At present comparative distinctions among teacher preparation programs in ELA are problematic.

The fourth and fifth columns, effects for grades 4-5 and 6-9, are provided to illustrate how the data might appear if results are subsequently reported in narrower grade bands that correspond to the State's certification structure. Louisiana currently offers certification in grades 1-5, 4-8, and 6-12, among others. These ranges generally correspond to the 4-5, 6-9, and overall data. One of the efforts for the future of this project is to examine analyses by specific certification programs such as the 1-5 or 4-8 certification programs. In cases where a university had less than 10 new program completers in a grade band, results are not reported.

Table 14: *Teacher Preparation Program Effects: Mathematics 2004-2005*

Level	University Teacher Preparation Program	Effect for Overall Achievement (CI)	Effect for Grades 4-6 (CI)	Effect for Grades 6-9 (CI)
III	Undergraduate Univ. B	0.7 (-2.3, 3.6)	0.7 (-3.3, 4.7)	1.3 (-2.7, 5.3)
III	Undergraduate Univ. E	-1.0 (-4.2, 2.1)	-3.6 (-8.3, 1.1)	2.4 (-0.9, 5.6)
III	Undergraduate Univ. H	0.2 (-4.4, 4.7)	1.5 (-6.1, 9.2)	-1.7 (-5.8, 2.3)
III	Undergraduate Univ. K	-0.1 (-4.7, 4.6)	-1.2 (-8.1, 5.7)	1.1 (-4.2, 6.3)
III	Undergraduate Univ. L	0.7 (-3.5, 4.8)	0.7 (-5.1, 6.5)	-0.2 (-4.8, 4.4)
III	Undergraduate Univ. M	-2.4 (-5.3, 0.6)	-4.5 (-8.4, -0.6)	0.4 (-2.8, 3.6)
III	Undergraduate Univ. N	-1.2 (-5.5, 3.2)	ISD	ISD
III	Alternative Cert. Univ. M	-0.3 (-4.1, 3.6)	ISD	ISD
III	Alternative Cert. Univ. P	0.3 (-3.1, 3.7)	1.6 (-3.8, 7.1)	0.0 (-4.1, 4.1)
II	Undergraduate Univ. A	-3.2 (-18.8, 12.4)	ISD	ISD
II	Undergraduate Univ. D	-2.5 (-9.8, 4.8)	-14.0 (-32.9, 5.0)	1.8 (-2.0, 5.6)
II	Undergraduate Univ. F	-2.5 (-5.4, 0.5)	-0.3 (-5.5, 4.8)	-4.2 (-7.1, -1.2)
II	Undergraduate Univ. G	-4.7 (-9.9, 0.4)	ISD	ISD
II	Undergraduate Univ. J	-5.0 (-12.8, 2.9)	-7.7 (-17.6, 2.2)	ISD
II	Alternative Cert. Univ. H	-3.8 (-10.5, 2.9)	ISD	ISD
II	Alternative Cert. Univ. Q	-3.0 (-7.7, 1.7)	ISD	ISD
I	Undergraduate Univ. I	-2.7 (-5.2, -0.2)	-3.0 (-6.7, 0.7)	-1.7 (-4.4, 1.1)
I	Alternative Cert. Univ. B	-8.4 (-13.6, -3.3)	ISD	ISD

Table 14 (continued)

I	Alternative Cert. Univ. G	-6.6 (-12.9, -0.3)	ISD	ISD
I	Alternative Cert. Univ. L	-5.7 (-10.0, -1.4)	ISD	-4.1 (-8.3, 0.0)

Note. The top number in each cell is the mean adjustment to student outcome that would be expected based upon a standard deviation of 50. The numbers in parentheses are the 95% confidence intervals. ISD indicates insufficient data in that cell to report results. The minimum number of new teachers per cell was set at 10.

Nine programs were identified whose effect was not statistically significantly different from experienced teachers and whose point estimate was relatively close to that of experienced teachers. The effects for Level III programs ranged from -2.4 to +0.7. One university (M) had both an undergraduate and an alternative certification program in Level III. Seven programs fell in Level II. All of these programs point estimates were below the point estimate for average first year teachers, but exhibited such great variability that their 95% confidence interval included the effect for experienced teachers. Four programs fell in Level I with alternative certification programs being over represented (3 of 4). Effects for Level I programs ranged from -2.7 to -8.4.

Generally the same cautions are warranted regarding comparisons among new teacher preparation programs in mathematics. The generally low level of precision in estimating their effects is suggested by the relatively large confidence intervals. In a few instances the 95% confidence intervals are not overlapping (e.g., Alternative Certification B versus Undergraduate B). The reliability of these few differences is questionable in a context including such a large number of comparisons.

Table 15: *Teacher Preparation Program Effects: Science 2004-2005*

Level	University Teacher Preparation Program	Effect for Overall Achievement (CI)	Effect for Grades 4-6 (CI)	Effect for Grades 6-9 (CI)
III	Undergraduate Univ. B	2.2 (-4.5, 8.9)	0.1 (-12.2, 12.4)	4.7 (2.2, 7.2)
III	Undergraduate Univ. C	0.8 (-6.0, 7.5)	2.7 (-12.5, 17.9)	-0.3 (-6.0, 5.4)
III	Undergraduate Univ. D	0.1 (-2.9, 3.0)	ISD	ISD
III	Undergraduate Univ. E	0.1 (-2.6, 2.8)	0.8 (-3.1, 4.8)	-0.6 (-4.2, 3.1)
III	Undergraduate Univ. F	0.1 (-4.4, 4.6)	-2.2 (-8.1, 3.7)	1.6 (-2.8, 6.0)

Table 15 (continued)

III	Undergraduate Univ. G	-0.6 (-4.4, 3.2)	ISD	ISD
III	Undergraduate Univ. H	-0.8 (-4.7, 3.2)	1.3 (-3.4, 6.0)	-2.6 (-7.3, 2.0)
III	Undergraduate Univ. I	-1.2 (-3.6, 1.2)	1.1 (-4.5, 6.7)	-1.1 (-3.3, 1.1)
III	Undergraduate Univ. J	-1.7 (-4.2, 0.8)	-1.1 (-5.3, 3.1)	-1.9 (-4.2, 0.4)
II	Undergraduate Univ. L	-2.0 (-5.1, 1.0)	-3.5 (-9.1, 2.2)	ISD
II	Undergraduate Univ. M	-2.2 (-6.6, 2.2)	-4.0 (-9.5, 1.6)	0.4 (-7.0, 7.9)
II	Alternative Cert. Univ. B	-2.3 (-8.6, 4.0)	ISD	ISD
II	Alternative Cert. Univ. L	-3.4 (-8.5, 1.7)	-4.6 (-10.8, 1.6)	ISD
II	Alternative Cert. Univ. P	-7.5 (-16.8, 1.7)	-6.1 (-17.0, 4.7)	-15.2 (-17.3, -13.1)
I	Alternative Cert. Univ. E	-2.5 (-5.0, -0.1)	ISD	ISD
I	Alternative Cert. Univ. G	-3.3 (-5.7, -0.9)	ISD	ISD
I	Alternative Cert. Univ. H	-3.3 (-6.4, -0.2)	ISD	ISD
I	Alternative Cert. Univ. M	-4.3 (-7.0, -1.5)	ISD	ISD

Note. The top number in each cell is the mean adjustment to student outcome that would be expected based upon a standard deviation of 50. The numbers in parentheses are the 95% confidence intervals. ISD indicates insufficient data in that cell to report results. The minimum number of new teachers per cell was set at 10.

Nine undergraduate preparation programs were identified that were not statistically significantly different from experienced teachers and fell within the mean standard error of experienced teachers (Level III) in science. The point estimates of university effects for science in Level III ranged from -1.7 to +2.2. It is also interesting to note that University B met the criteria for a Level IV designation in the upper grades (6-9). Five programs fell within the working Level II designation and four fell within Level I. Effects for Level I programs ranged from -2.5 to -4.3. All of the Level I programs were alternative certification programs. All of the confidence intervals for program effect were overlapping.

Table 16: *Teacher Preparation Program Effects: Social Studies 2004-2005*

Level	University Teacher Preparation Program	Effect for Overall Achievement (CI)	Effect for Grades 4-6 (CI)	Effect for Grades 6-9 (CI)
III	Undergraduate Univ. B	-0.3 (-2.6, 2.1)	0.5 (-3.9, 4.9)	-0.5 (-3.0, 2.0)
III	Undergraduate Univ. E	-1.8 (-4.8, 1.2)	-3.2 (-7.0, 0.6)	0.0 (-4.4, 4.3)
III	Undergraduate Univ. F	-1.4 (-4.6, 1.8)	-0.6 (-5.8, 4.5)	-2.1 (-5.0, 0.8)
III	Undergraduate Univ. G	-0.2 (-3.7, 3.2)	ISD	ISD
III	Undergraduate Univ. H	0.3 (-3.9, 4.5)	1.0 (-6.4, 8.5)	0.0 (-3.8, 3.8)
III	Undergraduate Univ. L	-1.5 (-5.0, 1.9)	-3.5 (-7.7, 0.8)	-0.5 (-5.9, 4.9)
III	Alternative Cert. Univ. B	-2.0 (-5.4, 1.3)	ISD	ISD
III	Alternative Cert. Univ. G	0.3 (-4.6, 5.3)	ISD	ISD
III	Alternative Cert. Univ. M	-0.9 (-7.9, 6.1)	ISD	-2.7 (-9.9, 4.5)
III	Alternative Cert. Univ. O	-1.8 (-10.1, 6.6)	ISD	ISD
III	Alternative Cert. Univ. P	-0.7 (-4.3, 2.8)	-2.9 (-7.8, 2.0)	1.4 (-1.8, 4.5)
II	Undergraduate Univ. I	-2.3 (-5.2, 0.6)	-2.4 (-6.6, 1.8)	-2.5 (-5.6, 0.7)
I	Undergraduate Univ. A	-9.2 (-17.7, -0.8)	ISD	ISD
I	Undergraduate Univ. J	-5.1 (-10.1, -0.2)	-8.3 (-17.1, 0.6)	-1.6 (-3.7, 0.5)
I	Undergraduate Univ. M	-4.7 (-7.2, -2.3)	-4.8 (-8.8, -0.8)	-4.4 (-7.6, -1.1)
I	Alternative Cert. Univ. L	-6.1 (-10.0, -2.1)	ISD	-5.0 (-9.2, -0.9)

Note. The top number in each cell is the mean adjustment to student outcome that would be expected based upon a standard deviation of 50. The numbers in parentheses are the 95% confidence intervals. ISD indicates insufficient data in that cell to report results. The minimum number of new teachers per cell was set at 10.

Level III programs included 6 undergraduate and 5 alternative certification programs. The teacher preparation effects ranged from -2.0 to +0.3 for Level III programs. One program fell within the Level II designation and four fell within Level I. Effects for Level I programs ranged from -3.4 to -7.0. Once again all confidence intervals were overlapping.

VIII. Effects of Different Normative Comparison Bases

All of the analyses reported above compared new teachers to experienced certified teachers in a statewide analysis controlling for prior achievement, student demographic variables, class composition variables, and including a school building effect. The concern has been raised that even with these rather extensive controls that comparison of teachers across different schools may be problematic to the extent that they may be teaching in very different sorts of contexts. The argument advanced is that teachers should only be compared to teachers within the same school or to teachers teaching in very similar schools.

A localized model for each university was developed that used the data from only the schools in which the graduates of that program taught. This had the advantage of providing a comparison sample in which the new graduates and the experienced teachers were teaching in comparable schools, because they were teaching in the same schools. It has the disadvantage of creating different normative comparison samples for each university. The university effect from the statewide analysis was then compared to the results from the local sample.

This analysis revealed that while local comparison samples typically yielded very similar results, that they can be substantively different for some universities. For mathematics the correlation between the estimated university using a local versus a statewide comparison was only $r = .91$. The magnitude of the differences was fairly large in a few instances in the context of the program effects ranging from -4.8 points on the LEAP scale to +1.7, despite the strong overall correlation. The mean effect (+0.5) suggested a very small positive effect on the estimates of university effects based upon using a local versus a statewide comparison.

For ELA the correlation was $r = .82$ between the statewide and local estimate. The range of differences were similar in magnitude to mathematics (-4.1 to +2.6). The overall mean difference between local versus a statewide comparison was small (-0.02).

These analyses clarify that the choice of the comparative sample can have an impact on the decisions reached regarding particular universities. Unfortunately, both approaches have merits. The local approach assures that the comparison group is in an equivalent context because they are in the same context. However, it has the disadvantage of setting localized norms for teacher performance that in turn create unequal standards for teacher preparation effectiveness across institutions. This issue is one that will require further examination.

VIII. Reliability of VAA Estimates

Analyses were conducted to examine the reliability of teacher effectiveness estimates across years at the level of individual teachers. These analyses were based upon data drawn from the pilot years, 2002-2003 and 2003-2004, as well as the current year. Data were limited to the eight school districts which participated in both pilot years. For both mathematics and ELA, reliability was estimated using those teachers who were included in any two adjacent years or all three years. For each content area, the model that was developed based upon the statewide data analyses described above was used to estimate teacher effects. For these analyses the teacher group membership variables were dropped because the focus was on estimating reliability at the individual teacher level. The empirical Bayes intercept residual was then obtained for each teacher for each year. This value was a measure of the degree to which the teacher's measured effectiveness differed from the effectiveness that would be predicted by all the variables in the model. Conceptually, this result could be described as an estimate of the unique teacher effectiveness plus measurement and model error. The current estimates of reliability are considered to likely be lower bound estimates for two reasons. First, the model is still being developed and further refinements may yield more precise estimates. Second, it is likely that more reliable estimates can be obtained by using multiyear averages.

It is also worth noting that these estimates are likely to underestimate the reliability of estimates of the university teacher preparation program's (TPP) efficacy. Conceptually, the individual teacher estimates might be described as the items lying within the scale of assessing TPP efficacy. Generally, items are substantially less reliable than scales. The reliability estimates for teacher level data across years, with different collections of students, and different forms of the assessments are presented below.

Table 17: *Correlation between Individual Level Teacher Efficacy Estimates Across Years*

Teacher group	English Language Arts	Mathematics
	(<i>n</i>)	(<i>n</i>)
All teachers	.46	.50
2003 with 2004	(342)	(359)
Teachers with 10 or more students	.58	.50
2003 with 2004	(218)	(244)
All teachers	.34	.32
2003 with 2005	(281)	(293)
Teachers with 10 or more students	.35	.33
2003 with 2005	(218)	(246)
All teachers	.48	.51
2003 with 2004	(355)	(373)
Teachers with 10 or more students	.48	.55
2003 with 2004	(218)	(246)

Generally these initial *individual* level reliability estimates are promising given that generalization across collections of students, 12 months, and tests are all being examined simultaneously. Composite program level reliability estimates and multiyear estimates would both be predicted to be more stable. Additionally, further refinement of the VAA model may yield additional increments in reliability. Based upon the absence of a substantive difference between the correlations for teachers with data for 10 students versus all teachers, adding this requirement does not appear warranted.

IX. Teacher ACT Scores at College Admission and Effectiveness

The research plan included examining the relationship between teachers' ACT scores prior to entering university education and their effectiveness. Initial review of the data revealed that ACT scores for new teachers were missing for 69% of new teachers. Preliminary analyses for ELA and mathematics did not reveal a statistically significant association between ACT scores and the estimate of individual teacher effectiveness that was used in the reliability analyses described above. However, no substantive conclusions should be drawn from this finding other than that the data available as of this date are too incomplete to provide clear information about this issue. At present the author does not know the extent to which the data are missing at random versus in a systematic manner. Additional research regarding this issue is needed.

X. Secondary Analysis Findings

Practitioner Teacher Effectiveness & New Teacher Assignment

In analyzing the data two phenomena of note emerged. The first phenomenon was that teachers teaching on a practitioner teacher's license were less effective than experienced teachers (mean effect for mathematics -3.1 and for ELA -5.1). These negative effects were typical of the less effective teacher preparation programs' effects in those content areas. It is important to note that while a teacher is in a practitioner teacher program that although they have not completed their training program, they are the teacher of record for students as they are completing their training.

What was more striking than the effectiveness of new teachers were the differences in teacher assignments based on their years of experience and license type. Table 18 below presents the selected mean demographic variables related to the characteristics of the students taught by experienced certified teachers, new teachers from undergraduate programs, new teachers from alternative certification programs, and practitioner teachers.

Table 18: *Student Demographic Factors by Assigned Teacher*

Content	Teacher	Prior Achievement	Percentage African American	Percentage Free/Reduced Lunch	Special Education
ELA	Experienced	+1.3	45%	61%	11%
	New – UG	-1.3	47%	64%	11%
	NEW – AC	-6.4	56%	67%	12%
	Practitioner	-16.4	69%	75%	15%
Math	Experienced	+2.0	44%	60%	11%
	New – UG	-1.1	48%	65%	11%
	NEW – AC	-6.6	60%	69%	11%
	Practitioner	-11.8	67%	72%	14%

Table note: Prior achievement is the difference from the State average prior achievement on the LEAP scale (Mean = 300, standard deviation = 50).

Interestingly although the differences in the student groups taught by new undergraduate program completers and experienced teachers consistently favor experienced teachers, the differences are fairly modest. In contrast the differences between the assignments of new alternative certification program completers and practitioner teachers versus experienced teachers might be described as dramatic. It is worth noting that these data clearly suggest that the lower achieving, poorer, disabled and minority students are much more likely to get a teacher whose preparation is not complete or was abbreviated than are their more advantaged peers. This issue obviously has policy implications that go beyond this technical report.

XI. Summary

Analyses were conducted to replicate and extend the pilot work to 66 of Louisiana's 68 school districts. The two school districts not represented are expected to be included in all future years' work. Their exclusion was the result of local issues in getting data ready for the statewide data system. 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 93.8% linkage rate for student data across years was very encouraging. Following the construction of the database, a number of analyses were conducted. First, preliminary statewide OLS regression analyses were conducted examining the use of various predictors of achievement. Second, mixed linear models of student achievement were developed for each content area using student level and classroom level variables to predict achievement. These models nested students within teachers and teachers within schools and included teacher and school effects. Third, these models were then applied in each content area to implement a VAA of teacher preparation. Fourth, the VAA analyses were followed by an examination of the reliability of teacher effect estimates over years at the level of individual teachers. Finally, the planned examination of the relationship between teachers' teaching efficacy and ACT scores was not appropriate due to the large amount of missing data and the lack of information about the reason those data were missing.

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 three other domains strengthened that relationship as did adding student level demographics. Adding a second year of achievement data strengthened the relationship to a greater degree than adding student demographic variables. School building level variables did not make a substantive contribution to variance accounted for.
2. The mixed linear models developed for each of the content areas shared a great deal in common. Although considerable overlap was present, some differences in models emerged at the classroom level. Once prior achievement in the four content areas was included in the model, the division of variance between students, teachers, and schools stabilized to a great extent, suggesting that these predictors accounted for a sizeable portion of the variance in achievement.
3. Examination of teacher effects by years suggest redefining new teachers as first and second year teachers and experienced teachers as teachers in their third year and beyond.
4. VAA of teacher preparation programs revealed that the precision of the estimates was somewhat problematic as suggested by typically relatively large confidence

intervals. These findings are both interesting and challenging in light of the degree variability in the point estimated effectiveness of programs. Some programs did yield results suggesting that their graduates were statistically significantly less effective than experienced teachers and these programs' point estimates fell below the mean effectiveness of new teachers in that content area. In a few instances confidence intervals for programs did not overlap within a content area, but this was an infrequent result. It was also interesting to note that in quite a number of instances the estimated teacher preparation program effect was exceedingly close to that of new teachers.

5. The reliability of individual level estimates of teacher effectiveness were somewhat promising, but certainly warrant additional research. It is important to note that these estimates represent a very severe standard in that they are separated by a year's time, are based upon different collections of students, and different tests. It is also worth noting that individual estimates are likely to severely underestimate the reliability of university level composites that are based upon many individuals.
6. The estimates of university effectiveness are somewhat sensitive to whether a local or a statewide model is applied. These models create different normative comparisons and contribute different information about the university preparation programs. This is a substantive issue that will require further examination.

These findings extend the findings of the two prior years' research. A number of issues remain if this sort of modeling were to be used to help universities assess their teacher preparation programs. First, a standard model will need to be developed and employed across years. Although a few difficult decisions may need to be made, on the whole, the degree of similarity in models across content areas suggests that this is achievable. Replication with an additional year's statewide data would contribute important information to this process.

A second issue that has arisen is concern regarding the accuracy of the student-course-teacher link in the statewide data system. Errors are likely more common for this year than any other since it is the first year of statewide implementation. Work is already underway to improve the system. Follow-up work will be needed examining the accuracy of the linkages. It is important to note that at the level of universities this is likely to be unsystematic error and as a result is likely to reduce the estimates of the impact of universities on student achievement. Reducing this source of error is likely to lead to more precise estimates.

Third and perhaps most importantly, additional research is needed examining methods of increasing the precision of the university effect estimates. Additional data will permit examination of how the amount of data influences estimates. Examination of alternative modeling strategies may also turn out to be helpful.

Major foci for additional planned research are summarized briefly below.

1. Estimates are needed of how many teachers are needed in order to report an estimate for a university. A study using sampling is planned to establish an estimate of the impact of varying number of teachers on the estimate of university effects.
2. The clustering of the graduates of particular teacher preparation programs within school districts present additional challenges. Although all of the universities studied have new graduates in multiple districts and in some cases do not exhibit substantial clustering, in some cases graduates clearly cluster within specific districts. The impact of clustering on analytic models warrants additional scrutiny.
3. The current models estimate the contribution of the current year's teacher to the students' total achievement for all school years. This underestimates the contribution of this year's teacher to the learning that occurs this year. Although Louisiana does not use the sort of vertically aligned tests that would contribute directly to this sort of analysis, further analytic work will examine whether it is possible to more closely approximate this type of analysis. Also recently reported findings regarding errors introduced by different content emphases in tests across years raises additional issues that warrant further examination in this regard (Martineau, 2006).
4. Although there were so many missing data for ACT scores that analyses based upon these data are not promising, the state has much more complete data for teacher certification tests. These are somewhat problematic conceptually because they happen at the end of teacher preparation; however it may be useful to examine them as a predictor of teacher performance.
5. The models presented herein represent a covariate approach to building the models in question. Some scholars in the area have argued that a repeated measures multivariate approach provides a stronger analytic approach while others have argued that at a practical level, data have not demonstrated the practical superiority of this more complex approach (Sanders & Horn, 1998; Tekwe et al., 2004). Although the curriculum data at a State level do not yet exist to fully implement a layered model, follow-up research examining the possibility of implementing such a model within the context of Louisiana's enormous data system will be examined.

In summary, despite some substantial changes in the data and more modest changes in the analytic strategy the models developed in this study were generally similar to the prior years' pilot work. The data suggest that some teacher preparation programs in Louisiana are producing graduates who are similar to existing teachers in their effectiveness as reflected by student performance on standardized achievement tests. They also suggest that it is possible to detect a minority of teacher preparation programs who are producing teachers who are less effective than experienced teachers. Although a

number of analytic issues remain to be resolved, no obvious barriers to completing the work needed to address those issues have emerged. Chief among the analytic issues is developing a clear understanding of the amount and type of data needed to develop useful information. While more interesting challenges still lie ahead, the data collection, data management, and analytic work is being completed that will set the occasion for addressing these new challenges. The most interesting challenge may be developing an approach to using the data that contributes to a continuous improvement and growth model rather than a political/professional conflict that distracts all from the important work at hand. The most interesting challenge once/if the technical issues are resolved is how to use the data to help those who prepare teachers provide to Louisiana's sons and daughters with more effective teachers than they have had in the past.

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.
- 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.
- McCaffrey, D. F., Lockwood, J. R., Korte, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND corporation.
- McCaffrey, D. F., Lockwood, J. R., Korte, 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. & Burns, J. L. (2006). Value added assessment of teacher preparation: An illustration of emerging technology. *Journal of Teacher Education, 57*, 37-50.

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

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.

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.

Sheldon, K. M. & Biddle, B. J. (1998). Standards, accountability, and school reform: Perils and pitfalls. *Teachers College Record, 100*, 164-180

Thompson, B. R. (2004). Equitable measurement of school effectiveness. *Urban Education, 39*, 200-229.

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.

Webster, W. J., & Mendro, R. L. (1997). The Dallas value-added accountability system. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 81-99). Thousand Oaks, CA: Corwin Press.