Technical Report of:

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

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

Acknowledgements

This report is based upon data provided by the Division of Planning, Analysis, and Information Resources (teacher and course data) and the Division of Student Standards and Assessments (student achievement data). The author would like to thank Michael Collier and Dr. Bobby Franklin for their assistance and guidance with matters related to linking students, courses, and teachers without which this work would not have been possible. The author would like to thank Dr. Fen Chou for her guidance regarding student level achievement data. The author would also like to thank Dr. Mary Helen McCoy for her assistance regarding interpreting certification data.

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 in part by award CT-04/05-VAM-EPE-01 from the Louisiana Board of Regents through the **Center for Innovative Teaching and Learning (CITAL).**

Table of Contents

Abstract

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

A preliminary set of analyses was conducted linking students to courses and courses to teachers based upon data collected by the Louisiana Department of Education's Divisions of Planning, Analysis, and Information Resources and Student Standards and Assessments. An analysis of covariance, a weighted analysis of covariance, and a hierarchical linear model (HLM) approach were examined across English-language arts and mathematics. These models examined changes in student achievement status nested within teachers' classrooms after controlling for prior achievement, demographic factors, and classroom context factors. These analyses were completed based upon the 10 parish school systems that participated in a pilot project collecting data regarding student course enrollment. Data from approximately 40,000 students in grades 4 through 9 and more than 1000 teachers in each content area from these parishes contributed to these analyses.

Results suggested that the strongest predictor of current achievement is prior achievement and that demographic factors are decreasingly important as more years of achievement data were available as predictors. Statistically significant differences were obtained and they typically, but not always, favored experienced teachers over new teachers. Based upon previous work in this area the effects found in this study are likely to be a lower bound for the magnitude of the effects. When data are available to link students to teachers across more than one school year it is anticipated that the size of the effects are likely to increase. The hierarchical linear models approach appears to be the most flexible and appropriate to this type of assessment. Previous research suggests that the current analyses may underestimate the magnitude of teacher effects. This suggests the need for additional longitudinal analyses that can match students to teachers across more than one school year to obtain a more accurate estimate of the size of these effects.

A number of issues remain to be resolved in future work. First, an a priori model for assessing teacher preparation programs may be desirable. Second, structures for integrating students enrolled in multiple courses in the same content need to be explored. This is a particularly pressing need for students in special education. Third, some additional investigation into the extent that students' assignment to teachers changes during the course of a year within schools appears to be needed to address a potential confound of the data. Fourth, all of the data examined herein were based upon relative comparisons within the State. An assessment program that can link State data to national benchmarks would be particularly useful. Finally, if a true statewide assessment system similar to this pilot were to be adopted, the practical considerations for data management, data analysis, and communication to stakeholders would be substantial.

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

I. Overview

Assessing the effectiveness of newly prepared teachers is a critical challenge confronting universities, school districts, the Board of Elementary and Secondary Education (BESE), and the Board of Regents (BoR). 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 metric, the extent of 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 hundreds of classrooms distributed across Louisiana 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.

Pragmatic Issues

One obvious means of addressing the pragmatic challenge of collecting a large amount of student achievement data is to use the achievement data that Louisiana currently collects. Broad spectrum standardized achievement data based upon measures that have established reliability and validity data are available for students from grade 3 through high school. A major barrier to attempts to use these data for this purpose is that no practical means has existed for linking students to teachers. However, with the development of the LEADS database by Louisiana's Department of Education's Division of Planning, Analysis, and Information Resources, that barrier will soon be overcome. Initial pilot data from two years are available to begin examining modeling options based on LEADS for assessing teacher effectiveness; however, these data are limited to 10 parishes and as a result do not provide a basis for a statewide assessment model. The planned implementation of LEADS on a statewide basis would solve one of the major practical barriers: linking students and teachers.

A second major pragmatic issue is the fact that the current comprehensive testing program only currently extends to grade 3. As a result, assessment of new teachers would be limited to grades 4 and beyond (to provide at least 1 year of pretest data). The State and universities have substantial interest in assessing the efficacy of teachers in the early grades (K-3). However, the adoption and planned statewide use of the *Dynamic Indicators of Early Basic Literacy Skills* (DIBELS) may provide a basis for examining teacher efficacy in **reading only** in grades K through 3. The DIBELS program will call for multiple assessments per year and may provide an important new element for assessing the efficacy of teacher preparation programs in the domain of early literacy. The implementation of DIBELS is at a sufficiently preliminary stage at this point that its inclusion in the current examination is premature.

Analytic Issues

Once data are available reflecting student achievement across years and matching those students to teachers, the minimum requirements for developing a value added assessment of teacher efficacy will have been met. However, the analytic issues underlying the assessment of teacher efficacy are formidable (see for example McCaffrey, Lockwood, Koretz, & Hamilton, 2003; Meyer, 1997; or Rowan, Correnti, & Miller, 2002). It is generally assumed and supported by previous research that student learning is moderated by contextual and individual differences in addition to teachers. As a result, two university programs that are equally effective would be expected to appear to differ in their effectiveness if they served groups of students and schools that differed substantially on variables that influence learning.

Several analytic models have been developed that attempt to control for individual differences and contextual differences to permit more fair comparisons. However, no single approach has yet become the accepted professional standard. The purpose of this project is to examine three analysis models, the information they return, and the degree to which they fit Louisiana's data. These analysis approaches are an Analysis of Covariance, a weighted Analysis of Covariance, and a Hierarchical Linear Model (HLM). Each of these approaches will be described below immediately prior to presenting the relevant data.

Known Limitations and Strengths of the Value Added Assessments

It is important to recognize several important limitations of value added assessments as applied to teacher effectiveness at the outset. First, regardless of the care taken in developing a value added assessment for teacher preparation, there will remain consumers who will object to the *concept* of assessing such a complex outcome as teacher preparation through statistical methods that require the data analyst to adopt one or more statistical assumptions (see Darling-Hammond, 1997). A related concern is that although every effort may be taken to use the best available data to remove the effects of variables such as poverty, it cannot be known whether the groups of teachers have truly been equated statistically on all important factors. Additionally, there will always remain some potentially important variables (e.g., parental level of education) for which data will not be available. It is worth recognizing that previous research using data that are longitudinal in nature and include multiple years of teacher data within powerful analytic designs have suggested that status (i.e., race) and context (i.e., percentage of poor students in class) may not be as important as has been thought (Ballou, Sanders, & Wright, in press; Wright, Horn, & Sanders, 1997). It appears that previous research may have exaggerated the importance of these variables due to the absence of an appropriate longitudinal framework for studying the data. Stated differently, including multiple years' achievement data for students across multiple teachers may tap into the same underlying variability in achievement as measuring race, poverty, and parental education.

A second limitation is that the strongest value added approaches are based on assessment of student learning. That is, a series of tests that are aligned with one another and are vertical in nature are given so that results one year are directly comparable to results the next year. This is not the case in Louisiana. The scaled score for the LEAP 21 has no directly comparable meaning in reference to the Iowa Tests of Basic Skills that will precede or follow it. Although comparisons can be made after an appropriate standardization of the scores within years, this is a weaker approach. The assessment of teacher effectiveness in this case will examine how much a teacher changed the students' status within the group of students rather than specifically how much information that teacher taught that student. Although it is generally learning that determines this status within the group, prior research suggests that assessments based upon status will underestimate the impact of teachers when compared to assessments that directly assess learning (see Rowan et al., 2002). Additionally, Rowan et al.'s work suggests that the type of analyses that are possible with only one year of teacher data may substantially underestimate the size of teacher effects on student achievement.

A third broad limitation is that the strongest value added approaches use a cross classified HLM or mixed model approach (see McCulloch & Searle, 2001, or Raudenbush & Bryk, 2002). However, at present only one statistical software package is available that will accomplish this (Educational Value Added Assessment System, EVAAS). This is a proprietary system that can only be accessed through a service contracted through SAS. As a result it could not be examined in the current pilot. Additionally, using a cross-classified system requires multiple years of data for teacher assignment and student outcomes. At present that link is only possible through LEADS for one year. However, very shortly it will be possible to add a second year to the data analysis. It is also the case that a commercially available software package is due to be released shortly that will accommodate cross classification within HLM.

A fourth limitation is that the student achievement data are likely to include relatively many missing cases as they are merged across multiple years whose impact on the results will remain unknown. A fifth limitation is that using a Spring to Spring assessment window means that student gains after the standardized assessment actually contribute to the assessment of the following year's teacher, rather than the teacher who taught the student after testing was completed. The severity of this limitation will depend upon the amount of learning that takes place following standardized testing and the extent to which students retain that learning one year subsequently. Interestingly, it has been argued that end of year to end of year assessments have some strengths in comparison to beginning of the year to end of year assessments (McCaffrey et al., 2003).

Despite these limitations it appears that exploration of potential Value Added Teacher Preparation Program Assessment (VATPPA) is worthwhile. The most salient argument in its favor is that Louisiana has a massive data base that may shed light on the effectiveness of teacher preparation programs that is not being utilized for this purpose. In short, the answers provided by a VATPPA may be imperfect, but at present the State does not have any comparable information. Additionally, once a sufficient database is available that multi-year longitudinal data analysis is possible for thousands of teachers **and** students Louisiana will have a vehicle for examining teacher preparation in a manner that has not previously been accomplished. As more longitudinal data are gathered, the State's ability to examine a variety of relationships that may be of interest will be enhanced.

The following pages describe the process that was followed to examine the alternative approaches for VATPPA and their outcomes.

II. Data Merging Process

The target year of teaching assessed was the 2002-2003 academic year as reflected in the Fall 2002 LEADS database and the spring 2003 administrations of the ITBS and LEAP 21. Initial work was undertaken to resolve apparently duplicate records and multiple partially complete records. The details of this process are available from the author. Following this work the ITBS and LEAP 21 data files were merged and a further round of duplication resolution was undertaken. At the end of this process the data set contained 498,613 records, each representing 1 student. *Z*-scores were then calculated based on the LEAP 21 and ITBS scaled scores for English Language Arts (ELA), Mathematics, Science, Social Studies, and Reading Total (ITBS) within each grade level for 2003.

Following this, the 2000, 2001, and 2002 datasets for ITBS and LEAP 21 were examined and the initial work to resolve issues of duplicate records and multiple partially complete records was again undertaken. Following this the standard scores were then derived in the same manner as the 2003 data.

Once this work was completed, 2003 testing records were matched with 2002, 2001, and 2000 records. All match procedures required at least 2 independent indicators that the record matched in order for records to be matched. Initially students were matched across years if their SSN and last name matched. To accommodate name changes, all cases that had not matched previously were then re-examined to determine if new matches would arise if students' SSN, gender, and date of birth were compared. Finally, to account for recording errors for the SSN, a final round of matching was conducted in which the student's last name, first name, date of birth, and gender were compared. In remaining students who resulted in a complete match based upon these 4 criteria were added to the longitudinal data set. Table presents the outcome of the merging process.

Table 1: *Raw Match Outcomes*

Year(s)	Cases in Merged file	Percentage of 2003 cases
2003	498,613	100%
2002-2003	372,518	74.7%
2001-2003	260,307	52.2%
2000-2003	238,412	47.8%

Table 1 exaggerates the degree of the failure to match for two reasons. First, it includes fourth grade students in the match with 2001; however, 2003 4th graders were in grade 2 for the 2001 assessment, so can not contribute to the match. In addition there was a sharp jump in the number of students in special education who participated in LEAP or ITBS in 2002 with an increasing trend in 2003. As a result it was not possible to match many of the special education students beyond 2002. If the change in participation of special education students continues in succeeding years, the percentage matches should continue to increase.

Table 2 presents the percentage matches for each year band including only those cases who were eligible to match from 2003 (ie., 4th grade and above for 2002, 5th and above for 2001, etc.).

Table 2: *Percentage Match for Students Whose Grade Level in 2003 Could Match*

Year(s)	Number of Cases	Percentage of eligible 2003 cases
2002-2003	368,548	84%
2001-2003	259,716	69%
2000-2003	237,538	73%

Given the inevitable realities of students' absences, spoiled tests, moving, and clerical errors this seems to be an encouraging level of matching.

III. Preliminary Analyses

Prior to pursuing examination of approaches to implementing a VATPPA with Louisiana's achievement data, a series of statewide ordinary least squares (OLS) regression analyses were conducted to examine general patterns in the data. The selected data for English Language Arts (ELA) and mathematics are presented below. The balance of this report will focus on modeling efforts for ELA and mathematics because of their status as the "high stakes" assessment areas within the State.

A series of regression analyses was conducted in which progressively more variables employed as predictors and the multiple correlation between achievement in 2003 and predictor variables was examined. Initially, 256,831 students who were in grades 4 through 9 in the spring of 2003, who took either the ITBS or LEAP 21, and were promoted at the end of the 2002 school year were identified as initially eligible for inclusion.

Table 3*: English-Language Arts Preliminary Statewide Regression Analyses*

Predictors	Multiple correlation	Number of Students
Z-score: ELA 2002	.718	249,076
Z-scores 2002 achievement	.752	247,406
Z-scores 2002 achievement Student demographic factors	.767	247,069
Z-scores 2002 achievement School demographic factors	.755	245,878
Z-scores 2001 & 2002 achievement	.788	170,573
Z-scores 2001 & 2002 achievement Student demographic factors	.796	170,346
Z-scores 2001 & 2002 achievement School demographic factors	.790	169,461
Z-scores 2000 - 2002 achievement	.804	119,175
Z-scores 2000 - 2002 achievement Student demographic factors	.810	119,036
Z-scores 2000 - 2002 achievement School demographic factors	.806	118,453

Table 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, Section 504 status, Title I reading 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 identified as disabled, percentage of students identified as gifted, and percentage of students identified as having limited English proficiency.

Table 4*: Mathematics Preliminary Statewide Regression Analyses*

Predictors	Multiple correlation	Number of Students
Z-score: Mathematics 2002	.773	249,153
Z-scores 2002 achievement	.791	247,393
Z-scores 2002 achievement Student demographic factors	.798	247,058
Z-scores 2002 achievement School demographic factors	.794	245,865
Z-scores 2001 & 2002 achievement	.827	170,566
Z-scores 2001 & 2002 achievement Student demographic factors	.829	170,343
Z-scores 2001 & 2002 achievement School demographic factors	.829	169,454
Z-scores 2000 - 2002 achievement	.836	119,179
Z-scores 2000 - 2002 achievement Student demographic factors	.837	119,043
Z-scores 2000 - 2002 achievement School demographic factors	.838	118,457

Table 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, Section 504 status, Title I mathematics 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 identified as disabled, percentage of students identified as gifted, and percentage of students identified as having a limited English proficiency.

The most striking outcome of the preliminary statewide regression analyses was the strong relationship between achievement across years and the modest contribution of either demographic or school context factors. It is also clear that as the years of available achievement data increase the contribution of demographic factors attenuates substantially.

IV. Linking Students and Teachers

Following preliminary linking of data and analyses, the student achievement data were linked with the data contained in the LEADS database to connect 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’s Department of Education’s Division of Planning, Analysis, and Information Resources were linked to LEADS and the longitudinal educational achievement database. These data permitted identification of new teachers. An initial pool of 40,697 students was identified who attended school within a parish that participated in the LEADS pilot project and who were in grades 4-9 in the spring of 2003. Approximately 90% of students were linked to courses within the LEADS database. Approximately 2% were also dropped because they changed schools within the LEADS database during the school year.

Course codes from LEADS 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 testing 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. LEADS course codes that could not reasonably be linked to a standardized test (e.g., Jazz ensemble) were dropped.

Once the longitudinal, teacher, LEADS, and school demographic databases had been linked, teachers were assigned to one of four categories based upon the following criteria.

Table 5: *Teacher Group Assignment*

Group	Criteria
New teachers	1. Less than 3 years teaching experience. 2. Holds a C or L1 certificate. 3. Received a university degree within 5 years of the start of school.
Emergency Certified Teachers	1. Teachers who are teaching on an emergency temporary authority.
Regularly Certified Teachers	1. Has 3 years or more teaching experience. 2. Holds an A, B, C, L1, L2, or L3 certificate.
Other	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.

V. Analysis Models

Three data analytic approaches were examined. These models were an Analysis of Covariance model (ANCOVA), a Weighted Analysis of Covariance model (W-ANCOVA), and a Hierarchical Linear Model (HLM). The general approach and results for each analysis for ELA and mathematics are presented below.

1. ANCOVA and Weighted ANCOVA

The first two models examined were modest variations on an analysis of covariance model in which students’ achievement scores were adjusted for prior years’ achievement and the student's demographic status (i.e., minority, free/reduced lunch, gifted, special education, Title I eligibility, and limited English proficiency, gender, Section 504 status). This approach was completed in three analytic stages. In the first stage, the analysis was completed for all students for whom data were available for all three prior years and who had been promoted each of those years. Following this, the analysis was repeated with those students who had data for the preceding two years and who were promoted each of those years. Finally, the analysis was completed for those students who had valid data for just one year. These data were then combined for each teacher group using a weighted average for each result. The weight was based upon the number of cases contributing to that data analysis stage.

A second version of the ANCOVA model was also examined in which a weight was assigned to each student score that was the inverse of the number of students taught by that teacher. This is the Weighted ANCOVA model (W-ANCOVA). The W-ANCOVA model was implemented to assign weights to the student data so that all teachers contributed equally to the final result even though they taught different numbers of students. For example, in the unweighted model, if a seventh grade teacher taught five sections of 25 students, that teacher would be weighted five times as heavily as a fourth grade teacher who taught one class of 25 students. The W-ANCOVA model corrects for this by assigning the inverse of the number of students taught by a teacher as a weight to each student score so that all teachers can be weighted equally in the analysis.

The following assumptions of ANCOVA were examined and found to be tenable: normal distribution of error (histogram of standardized residuals), homoscedasticity (standardized scatter plot), linearity (bi-variate plots), multicollinearity (variance inflation factor), and independence (Durbin-Watson). The homogeneity of variances (Levene's test) was found to be tenable when tested for the un-weighted ANCOVA analysis for the complete sample for mathematics and ELA. The homogeneity of variances assumption was not tenable for the W-ANCOVA analyses. However, given that ANCOVA is relatively robust to violations of homogeneity of variances in the instance of large numbers of participants and that both the weighted and non-weighted solution would be examined the decision was made to move forward with the analysis.

The only assumption that appeared to be violated for both models was the assumption of homogeneity of the hyperplanes (assessed by *F* interaction term). However, given the sample size, even a very small effect that is based upon the entire sample is likely to be statistically significant. As a result the variance added to the model was examined and it was revealed that the test for heterogeneity of the hyperplanes only added .001 to .002 to the variance accounted for by the model following procedures recommended by Stevens (1996). This suggested that any violation is likely to be of a relatively small magnitude.

Table 6 presents the results of the ANCOVA and W-ANCOVA analyses for ELA and Table 7 presents the results for mathematics. Data were included separately for each university that provided at least 10 teachers and 100 students to the analysis. All remaining new teachers were collapsed into the other new teacher group. University preparation programs are identified by letter rather than by name because these are pilot data from the parishes that participated in the LEADS project rather than a sample from the State. It is virtually assured that the data from these parishes is a biased estimate of the effectiveness of these specific university preparation programs because data from approximately 90% of the state were unavailable and this occurred in a systematic manner. Only 10 parishes that volunteered to participate in the LEADS pilot participated. As a result, these data should be regarded as an examination of models for assessing teacher preparation programs, rather than an assessment of any specific program.

The data presented in Tables 6 and 7 are the adjustment to the mean outcome that would be expected for each group scaled to a standard deviation of 100 (the approximate standard deviation of the LEAP 21 which varies by grade and content area). The difference between adjusted means would be the expected difference in a student's score on the LEAP 21 if they were taught by the average teacher from the respective groups. So, if based upon demographic factors and prior achievement, a student's expected score on the LEAP 21 was 250, and that student was taught by a regular certified teacher, that student would be predicted to score 254.5 based on the ANCOVA model. In contrast, if that same student were taught by a new teacher from University B, the student would be expected to score 240.3 (ANCOVA model).

Table 6: *Outcomes for ANCOVA Models for English-Language Arts*

Teacher group	**ANCOVA Model**	**Weighted ANCOVA Model**
Regular certified	4.5 (3.2, 5.9)	3.5 (2.2, 4.9)
New University A	-6.8 (-12.3, -1.2)	-6.4 (-11.6, -1.2)
New University B	-9.7 (-15.9, -3.5)	-12.2 (-18.4, -5.9)
New: other universities	-8.1 (-16.7, 0.4)	-9.0 (-18.3, 0.3)
Emergency Certified	5.4 (-0.2, 11.1)	-0.1 (-5.8, 5.5)
Other	-3.0 (-7.0, 1.0)	-3.0 (-6.8, 0.9)

Table notes:

1. The top number in each cell is the mean adjustment to student outcome that would be expected based upon a standard deviation of 100. The numbers in parentheses are the 95% confidence interval.
2. The academic covariates were prior years' achievement in ELA, mathematics, science, and social studies for 1 to 3 years based upon the available data. Demographic covariates were minority status, free/reduced lunch, gifted, special education, Title I eligibility, gender, Section 504 status, and limited English proficiency.
3. The weighted ANCOVA weighted students' data to the inverse of the number of students that their teacher taught. This permitted all teachers to be weighted equally in the analysis.

Table 7: *Outcomes for ANCOVA Models for Mathematics*

Teacher group	ANCOVA Model	Weighted ANCOVA Model
Regular certified	5.7 (4.5, 6.9)	8.8 (7.6, 10.0)
New University A	-0.8 (-6.5, 4.9)	3.2 (-1.8, 8.2)
New University B	-8.0 (-14.1, -1.8)	-6.5 (-13.6, 0.6)
New University C	9.7 (-.02, 19.7)	14.8 (4.6, 25.0)
New: other universities	-6.8 (-18.0, 4.4)	-0.7 (-10.3, 8.9)
Emergency Certified	-1.0 (-6.8, 4.9)	5.2 (0.0, 10.4)
Other	5.9 (2.8, 9.0)	9.8 (6.8, 12.8)

Table notes:

1. The top number in each cell is the mean adjustment to student outcome that would be expected based upon a standard deviation of 100. The numbers in parentheses are the 95% confidence interval.
2. The academic covariates were prior years' achievement in ELA, mathematics, science, and social studies for 1 to 3 years based upon the available data. Demographic covariates were minority status, free/reduced lunch, gifted, special education, Title I eligibility, gender, Section 504 status, and limited English proficiency.
3. The weighted ANCOVA weighted students' data to the inverse of the number of students that their teacher taught. This permitted all teachers to be weighted equally in the analysis.

Across Tables 6 and 7 differences emerged between experienced regularly certified teachers and new teachers from differing universities. In most comparisons the difference favored the experienced teachers; however, in the case of University C in mathematics new teachers had a higher adjusted mean than experienced teachers. However, the 95% confidence intervals for new teachers from University C and experienced teachers overlapped.

Examining the pattern of results for the ANCOVA versus the W-ANCOVA, the ordering of effectiveness among experienced certified teachers and new teachers across ELA and mathematics was identical. The magnitude of the difference did change somewhat, but the differences were typically modest. Given that the focus of this analysis was on teacher effectiveness, a reasoned argument can be made that between the weighted and un-weighted solutions, the weighted is preferred.

Weighting did influence the results for emergency certified teachers; however, the pattern differed across content areas. Additionally, in some analyses emergency certified teachers appeared to be similar to experienced regularly certified teachers in effectiveness and in other instances they appeared less so. This is an issue that may warrant further investigation if this sort of approach is adopted.

2. Hierarchical Linear Models

The final analytic approach examined was hierarchical linear models (HLM). 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 for instability of estimates. Finally, they provide a framework in which the effects of multiple teachers across multiple years can be estimated and teacher effects across multiple groups of students over multiple years can be collapsed to a single estimate. This final advantage is not yet included in these data as multiple years of data linking students, course, and teachers with achievement outcomes was not available at the time these analyses were completed.

Building the current models. Analysis for both mathematics and ELA began by fitting an unconditional model and one with the prior year's achievement in mathematics, ELA, science, and social studies as predictors. In each case all of the prior year achievement scores exhibited statistically significant fixed effects and were retained. The random effects for achievement areas other than ELA were not significant in the ELA analysis and were set to 0. For mathematics the random effects other than mathematics were statistically significant, however their reliabilities were quite low (.05 to .12). Based in part on the recommendations of Raudenbush and Byrk (2002) and the desire to devise a comprehensible model, the contribution of prior achievement in domains other than mathematics was set as a fixed effect.

In the next stage demographic co-variates were entered as a block and set as random effects. Due to instability this model did not resolve at 10,000 iterations. Based on preliminary analyses and the conceptual questions the model was then re-specified with free/reduced price lunch, minority status, special education status, and Section 504 status set as random. For both ELA and mathematics the model then stabilized, but suggested that free/reduced price lunch was not a statistically significant random effect in the presence of minority status, special education status, and Section 504 status. In addition, gender was not a significant fixed effect for mathematics. The random effect for free/reduced price lunch was dropped from both models and the fixed effect for gender was dropped from the model for mathematics. Following this the random effect for each of the remaining demographic covariates was tested individually. Gender for ELA was the only remaining covariate for which the data suggested a random effect and it was so specified. Next, the effect of a series of classroom level covariates was tested, such as class size, and those that were significant were retained. Covariates were entered in the order suggested by prior *t*. Based upon results of a significant χ^2 for heterogeneity of student level variances, heterogeneity of student level variance was modeled based upon student gender. Gender was selected based upon a series of tests and provided the best fit to the data.

Codes for each of the teacher groups were then entered for the intercept effect at the teacher level of the model. This essentially modeled the effect of teachers being new teachers from particular universities, experienced regularly certified teachers, or some other designation on students' final level of achievement. Teacher group codes were then entered for the influence of prior achievement on students' achievement at the end of the year. If this effect were significant it would suggest that teachers in that group tended to either increase or decrease achievement disparity among students based upon prior achievement. The direction of the effect would determine whether the effect was to increase or decrease disparity. If a significant effect occurred in this block it was retained in the final model. Finally, the effect for teacher group was then modeled upon the varying status variables of minority status, special education status, Section 504 status, and gender (for ELA). This analysis examined the extent to which teachers in particular statuses were uniquely effective with students in these groups, above and beyond their general effectiveness that was already entered into the model. None of these effects was significant so none of them was retained.

The final models are presented below followed by the results.

Table 8: *HLM Model for ELA Achievement*

Model Level	**Variables Entered**
Student level covariates	Free/reduced price lunch *Minority status* Gifted *Special Education* Title I Reading eligibility Limited English proficiency *Gender* *Section 504 Status* *Prior Year ELA test result* Prior year test results: Science, Social Studies, Mathematics
Classroom covariates	Students' mean prior year achievement in ELA Percentage of students identified as gifted Percentage of female students
Classroom main effects	Codes for teacher group membership (see results below)

Table note: The effect of italicized student level covariate variables was modeled as varying across classrooms. All other student level covariates were set as fixed.

Table 9: *Outcomes for HLM Model for English-Language Arts*

Teacher group	**Effect for Overall Achievement** In comparison to experienced certified teachers (intercept)
New University A	-11.4 (-18.3, -4.4)
New University B	-15.5 (-23.0, -8.0)
New: other universities	-14.9 (-26.4, -3.5)
Emergency Certified	-2.1 (-9.7, 5.5)
Other	-7.6 (-13.0, -2.2)

Table notes:

1. All differences from regularly certified experience teachers were statistically significant at $\alpha < 0.05$.
2. The top number in each cell is the mean adjustment to student outcome that would be expected based upon a standard deviation of 100. The numbers in parentheses are the 95% confidence interval.

Based upon the HLM results, teachers with 3 or more years experience holding a regular teaching certificate (L1, L2, L3, A, B, or C) were more effective than new teachers from either University A or B or the collection of other new teachers. Overall, these results are quite similar to the results obtained through the ANCOVA analyses.

Table 10: *HLM Model for Mathematics Achievement*

Model Level	**Variables Entered**
Student level covariates	Free/reduced price lunch *Minority status* Gifted *Special Education* Title I Reading eligibility Limited English proficiency *Section 504 Status* *Prior Year Mathematics test result* Prior year test results: Science, Social Studies, ELA
Classroom covariates	Teacher's mean class enrollment Students' mean prior year achievement: mathematics Percentage of receiving free/reduced lunch Percentage of female students
Classroom main effects	Codes for teacher group membership (see results below)
Teacher effects moderating effect of prior achievement	As in classroom main effects

Table note: The effect of italicized student level covariate variables was modeled as varying across classrooms. All other student level covariates were set as fixed.

Table 11: *Outcomes for HLM Model for Mathematics Arts*

Teacher group	**Effect for Overall Achievement** In comparison to experienced certified teachers (intercept)
New University A	-2.7 (-9.4, 4)
New University B	-7.3 (-15, 0.3)
New University C	4.7 (-9.1, 18.5)
New: other universities	-6.2 (-15.2, 2.7)
Emergency Certified	-2.9 (-11.6, 5.7)
Other	-0.2 (-5.1, 4.8)

Table notes:

1. None of the differences from regularly certified experience teachers was statistically significant at $\alpha < 0.05$.
2. The top number in each cell is the mean adjustment to student outcome that would be expected based upon a standard deviation of 100. The numbers in parentheses are the 95% confidence interval.

A statistically significant result was obtained for the moderating relationship between teacher status for new teachers from University C and level of prior achievement. What this effect demonstrated was that new teachers attenuated the relationship between prior achievement and final outcome for their students by approximately 19%. Stated differently, recognizing the generally positive achievement outcome for new University C teachers it suggest that they were *relatively* more effective in producing similar results for all students regardless of prior achievement than were experienced regularly certified teachers.

As with the ELA results, the HLM mathematics results are quite similar to the ANCOVA models. They suggest that there are mean differences between new teachers in these parishes and experienced certified teachers, but they are not sufficiently reliable that they are statistically significant. The effect for University B is clearly on the borderline in each case.

Interestingly, the HLM approach uncovered an effect that could not be detected in the ANCOVA approach. This might be called a fairness or gap narrowing effect, in which final effects were more similar across students of different prior achievement than occurred in experienced regularly certified teacher's classrooms. If effects such as these were to emerge for special education or minority students they would be very encouraging indeed.

All HLM analyses were run with all new teachers and then again excluding new teachers who were teaching outside their area of certification. No substantive difference was obtained between the two approaches.

Summary

A series of exploratory analyses were completed to examine the feasibility of using the State's educational assessment data in concert with the LEADS database and other associated databases to assess teacher preparation programs. The degree of matching across years and the degree of matching between the LEADS data and the achievement data suggest that this approach is viable. The following points are primary findings of the data analyses.

- Generally the strongest relationships to achievement were with prior achievement.
- As the number of years of achievement data increased the contribution of demographic factors rapidly decreased to low levels.
- Statistically significant differences were obtained between new teachers and experienced certified teachers for student outcomes after controlling for prior achievement, demographic variables, and classroom context variables.
- Differences between new and experienced teachers generally favored experienced teachers, but that was not true in all cases.
- The HLM analysis detected an effect of new teachers from one university on the relationship between prior achievement and outcome that the ANCOVA model could not detect.

Based upon these analyses it would appear that it is indeed possible to use Louisiana's achievement and educational personnel databases to assess the effectiveness of teacher preparation programs. Using data across multiple years within a comprehensive Louisiana database would provide a basis for assessing all teacher preparation programs on the basis of the impact of their graduates on the students they teach. Although differences between the models assessed were modest, results generally suggest that the HLM approach is to be favored. The HLM approach is more flexible, can assess dimensions of effectiveness not assessed within the ANCOVA approach, better matches the natural structure of the data, and ties into the most powerful analytic approaches for longitudinal achievement data.

A number of issues remain if this sort of modeling is to be adopted as a routine form of assessment. First, it is likely that a more standardized model will need to be employed across years and content rather than adapting the model to each year and content area as was done herein. However, results were similar enough across both content areas that this approach should not be too problematic.

A second issue that arose is how to incorporate data from students who are enrolled in multiple courses in the same content area (e.g., two mathematics courses).

Although that was not common the issue did become apparent. The EVAAS system assigns the student to both courses and assumes that both courses contribute equally to the student's achievement. That approach obviously raises some a host of issues. For the present analyses, students with multiple courses in one content area were excluded. The majority of these cases were students in special education. A key limitation of the current data is that special education classes within the LEADS database do not current identify what the curricular content special education classes are intended to target. Adding content specific codes to the special education codes would permit a greater proportion of students to contribute to the analyses. It would also permit more accurate exploration of approaches to assessing the effect of teachers on students in special education.

One limitation of the current analyses that these data cannot address is the degree of class switching that occurs within schools during the year. All of the students who contributed to these analyses were in the same school for the spring LEAP 21/ITBS assessment as they were in the fall. However, we don't know how many of the 8th or 9th grade students had two different math courses, unless that plan was recorded when the LEADS data were completed. It is also the case that reassignments with that might occur between two 4th grade classrooms would not be captured. Additional research into the degree to which students are moved between classes or have multiple different courses within a content area, within a school year, within different grade levels in Louisiana, should be explored.

An additional limitation of these analyses is that all of the comparisons are relative to teachers within the State. If one of Louisiana's goals is to be more nationally competitive in the quality of the education provided to its sons and daughters, an out-of-state benchmark would be helpful. Further work using the national ITBS normative database as an out-of-state referent may prove useful in this regard.

A final issue to be resolved in the future is the relative utility of the types of models examined herein in contrast to the presumptively more powerful cross-classified mixed models approach. At the time these analyses were completed the data were not available to examine this approach. Additionally, no commercially available software is available that will implement cross-classified models. However, Scientific Software International Inc. has announced the release of commercial software that will permit this type of analysis. It will be possible to use in-State resources to examine this type of model in the near future. However, the option also exists to contract this type of evaluation through the SAS Corporation based upon EVAAS. This has the advantage that it is a well established and thoroughly tested system. EVAAS is also likely to be quite costly. If policy makers do decide to implement this type of evaluation system whether it is done within the State or through EVAAS, a substantial commitment to information technology, data management, data analysis, and communication to consumers will be required.

References

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

Darling-Hammond, L. (1997). Toward what end? The evaluation of student learning for the improvement of teaching. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 248-263). Thousand Oaks, CA: Corwin Press.

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.

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

Meyer, R. H. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review, 16,* 283-301.

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

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., Saxton, A. M., & Horn, S. P. (1997). The Tennessee value-added assessment system: A quantitative, outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press.

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.

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for evaluation. *Journal of Personnel Evaluation in Education, 11,* 57-67.