Digest of:

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

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

This work was supported 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

Acknowledgement

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.

Statement Regarding Technical Data

This report was prepared with a reduced emphasis on technical and statistical copy detail. A separate report is available that provides a much higher level of technical detail.

Abstract

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

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 how changes in student achievement status were related to teachers while taking other important non-teacher factors into account. These factors included students' 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 prior achievement data were considered. Teacher effectiveness generally favored experienced teachers over new teachers, but this was not always the case. Based upon prior research, it is anticipated that these analyses may be an underestimate of the size of the teacher's impact. Data linking students across multiple teachers and years will be needed to obtain a more accurate estimate.

A number of issues remain to be resolved in future work. First, an approach for handling students enrolled in multiple courses in one content area will need to be developed. Second, some investigation into how often students' assignment to teachers within schools changes during the course of a year is needed. Third, a model for benchmarking these results against a national rather than a Louisiana comparison would be helpful. 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. Assembling the Data

The target year of teaching assessed was the 2002-2003 academic year. The initial dataset included achievement testing data for approximately 500,000 students. These data were combined with prior years' data to create a longitudinal record of test results for each student for whom data were available. All test results were standardized within each year and grade level to permit comparisons. The record for each student also included demographic factors such as race, free lunch status, gender, and special education status.

The student academic data were then combined with data from additional databases that described students' assignment to teachers in the 10 parishes that participated in the LEADS project. This included approximately 40,000 students. Additional information from several sources was then added to the dataset that included information about teachers, classrooms, and schools. All teachers were then assigned to one of the following groups based upon the criteria in Table 1.

Table 1: *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.

III. Outcome of the Data Analyses

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). All three models produced similar results suggesting that the outcome of the assessment is not dependent upon a specific statistical approach. Because of its flexibility the author recommended the use of the HLM approach in the future. The following summarizes the outcome of the HLM analyses. The full results are available in the separate and more comprehensive technical report.

In order for a University's outcome to be examined it had to include at least 10 teachers and 100 students in the LEADS pilot data. 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

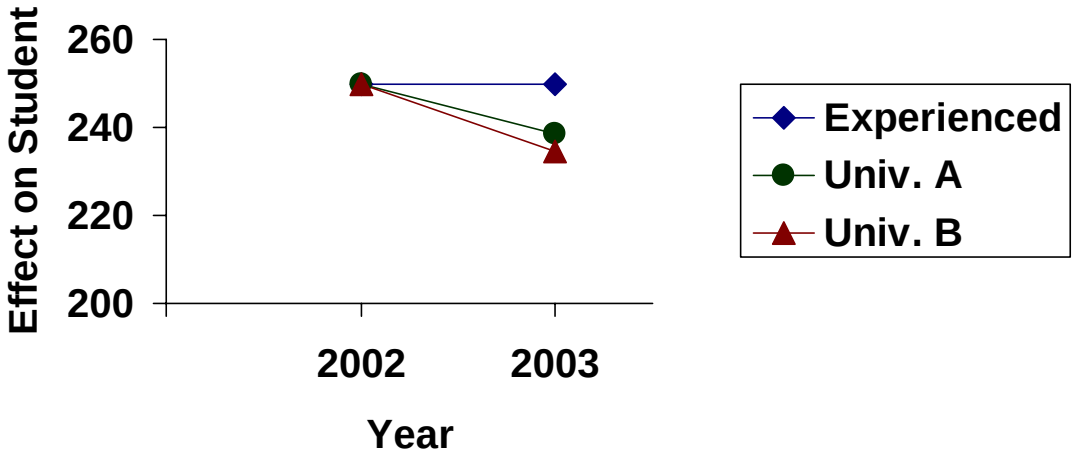
assured that the data from these parishes is a biased estimate of the effectiveness of these specific university preparation programs because data from approximately 90% of the state were unavailable and this occurred in a systematic manner.

Table 2 below describes the final HLM model for English-language arts that was derived. At the student level a large number of variables were used to predict what each student's achievement was likely to be for 2003. These included variables such as prior achievement in all four main content areas, minority status, and identification as gifted. These predictions were then adjusted based upon several classroom contextual factors. Finally, once all of these factors were taken into consideration the effect of the teacher was examined.

Table 2: *HLM Model for ELA Achievement*

Model Level	**Variables Used to Predict Achievement**
Student level predictors	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 adjustment factors	Students' mean prior year achievement in ELA Percentage of students identified as gifted Percentage of female students
Teacher effects	Teacher group membership (see results below)

Figure 1: ELA Teacher Effects
(LEAP Scale)



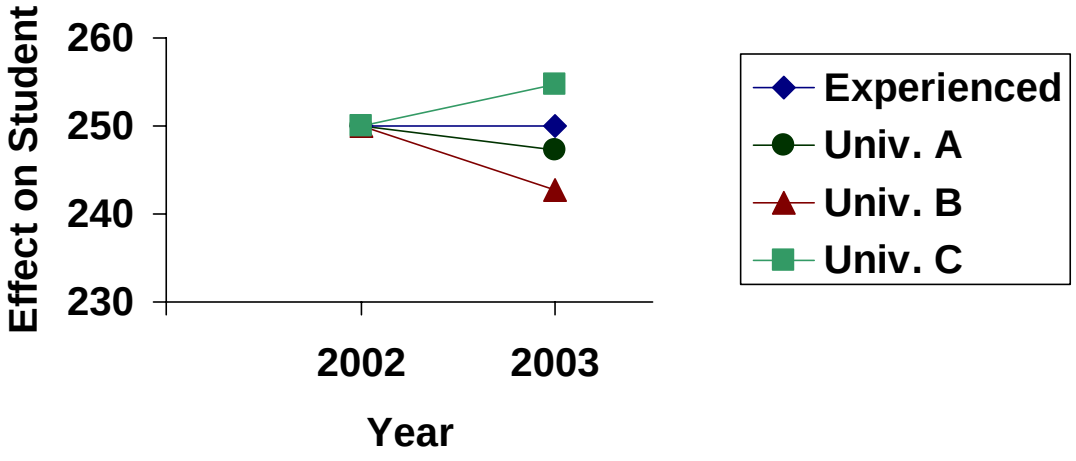
For these analyses experienced certified teachers were used as the benchmark. What Figure 1 shows is that as compared to students taught by experienced certified teachers a student taught by a new teacher from University A or University B would be expected to score approximately 11 (A) or 16 (B) points lower on the LEAP after considering the 12 student level predictors of achievement and the three classroom level adjustment factors.

Table 3 below presents the model for mathematics.

Table 3: *HLM Model for Mathematics Achievement*

Model Level	**Variables Used to Predict Achievement**
Student level predictors	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 adjustment factors	Teacher's mean class enrollment Students' mean prior year achievement: mathematics Percentage of receiving free/reduced lunch Percentage of female students
Teacher effects	Codes for teacher group membership (see results below)

Figure 2: Mathematics Teacher Effects
(LEAP Scale)



For these analyses experienced certified teachers were used as the benchmark. What Figure 2 shows is that as compared to students taught by experienced certified teachers a student taught by a new teacher from University A or University B would be expected to score approximately 3 (A) or 7 (B) points lower on the LEAP after considering the 12 student level predictors of achievement and the three classroom level adjustment factors. In contrast students taught by new teachers from University C would be expected to score approximately 5 points ***higher*** than students taught by experienced certified teachers after considering student background and the classroom context.

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, a decision will need to be made regarding how to handle data for students who are enrolled in multiple courses in the same content area (e.g., two mathematics courses). Second, the extent to which students move between teachers within the same school during the year needs to be examined to determine the extent to which that would confound results.

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.

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.