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

June 24, 2005

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

Cover Page 1

Contents 2

Abstract 3

Introduction 4

Data Merging Process 5

Preliminary Analyses 6

Linking Students and Teachers 8

Value Added Assessment: Hierarchical Linear Model 9

Reliability of VAA Estimates 14

Summary 15

References 17

Abstract

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

Analyses were conducted replicating pilot work examining the feasibility of using the Louisiana's educational assessment data in concert with the Louisiana Educational Assessment Data System (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. Although there were some differences, the models were strikingly similar across years. The reliability of individual level estimates of teacher efficacy across 12 months, different student groups, and different test forms were promising. As the number of years of achievement data increased, the contribution of demographic factors rapidly decreased to low levels. Some statistically significant differences were obtained between new teacher groups and experienced certified teachers for student outcomes after controlling for prior achievement, demographic variables, and classroom context variables.

Although the results of the current work are promising, a number of issues remain to be resolved in future work. First, an *a priori* model for assessing teacher preparation programs would be desirable for ongoing standardized assessment of teacher preparation programs. Second, structures for integrating student data across multiple courses across years need to be examined. Third, investigation into the extent that students' assignment to teachers changes during the course of a year within schools is needed to address a potential confound. Fourth, all of the data examined herein were based upon relative comparisons within the State. An assessment program to which can link State data to national benchmarks would be particularly useful. Finally, if a 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. Most importantly, resolving how to act on the assessment results would be the most crucial issue in determining the utility of the assessment program.

**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.

Year Pilot

Prior pilot work was completed examining three analytic models based upon Louisiana's educational data. The prior pilot analyses resulted in a number of general findings relevant to year 2 of the pilot work. First, the year to year association between the educational assessments used in Louisiana is sufficiently strong that the creation of a longitudinal analysis model based upon these data may be appropriate. Second, for those parishes pilot testing the LEADS data system, a sufficient quantity and quality of data are available to support such an undertaking. It is important to note that the LEADS data system was implemented statewide for the first time in 2004-2005. Third, all three analytic approaches examined suggested similar conclusions. However, a recommendation was provided to pursue hierarchical linear or linear mixed models based upon their flexibility and their potential for providing the most powerful analytic tools that are appropriate to the problem at hand. Fourth, using even the limited pilot data, some statistically significant differences were found between new teachers from specific universities' preparation programs and experienced certified teachers on teaching effectiveness. In some instances, teachers from some programs did not differ from experienced teachers at a statistically significant level.

A discussion of pragmatic issues, analytic issues, limitations of value added assessments (VAA), and strengths of VAA are provided in the year 1 report. This report describes the results of year 2 of the pilot research. Year 2 of the pilot research had three primary goals. The first goal was to replicate year 1 and examine the extent to which results would be comparable or disparate. The second goal was to examine the extent to which estimates of teacher efficacy are reliable from one year to the next based on the VAA being explored. It is important to note that *a priori* the expectations for goal two were very modest. Prior work generally has suggested the use of multiyear running averages to assess teacher efficacy (Sanders, Saxon, & Horn, 1997). The most powerful models for conducting this type of VAA are purported to be cross-classified models

(McCaffrey, Lockwood, Koretz, & Hamilton, 2003; Rown, Correnti, & Miller, 2002). The pilot data do not yet provide enough data to examine the reliability of cross-classified assessments. Finally, the pilot examined the technical and practical feasibility of conducting cross-classified analyses using recently produced commercially available software. In examining this third goal a number of issues arose related to the structure of the data, the specification of an appropriate model, and the capacity of the software that have resulted in the conclusion that it may not be possible to create a full cross-classified model with current commercially available software. However, the author has communicated with primary programmer for HLM 6 (Scientific Software Inc.) and an additional module is under development that may be available in the near term and may match Louisiana's analytic needs more precisely. Additionally, ongoing work is being conducted to examine modeling options/needs and available software.

This report describes the results of the replication of year 1 analyses and reliability analysis.

II. Data Merging Process

The target years of teaching assessed were the 2002-2003 and 2003-2004 academic years as reflected in the Fall 2002 and 2003 LEADS databases and the spring 2003 and 2004 administrations of the ITBS and LEAP 21. Pilot VAA of teacher preparation efficacy was previously reported for 2002-2003 and will not be repeated here. This report will describe the results of the replication for 2003-2004 and the reliability analysis that linked 2002-2003 with 2003-2004. Initial work was undertaken to resolve duplicate records and multiple partially complete records that described the same student. 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 486,157 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 2004.

Following this, the 2001, 2002, and 2003 datasets for ITBS and LEAP 21 were examined and the initial work to resolve issues of duplicate records and partially complete records was again undertaken. Following this, the standard scores were then derived in the same manner as the 2004 data.

Once this work was completed, 2004 testing records were matched with 2003, 2002, and 2001 records. All match procedures required at least 2 independent indicators 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 1 presents the outcome of the merging process. It is important to note that the structure of Louisiana's assessment program creates a natural attrition across years.

Students who were in third grade in 2004 cannot match with 2003, because no second grade assessment is available. Table 1 presents the total matches by year with 2004 and the percentage matches for each year band including only those cases which were eligible to match from 2003 (ie., 4th grade and above for 2002, 5th and above for 2001, etc.).

Table 1: *Percentage Match for Students Whose Grade Level in 2003 Could Match*

Year(s)	Number of Cases	Percentage of eligible 2003 cases
2003-2004	402,039	93.5%
2002-2004	298,117	80.4%
2001-2004	230,434	72.7%

Given the realities of students' moving out of state, absences, spoiled tests, and clerical errors this is a very encouraging level of matching.

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. 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 were employed as predictors and the multiple correlation between achievement in 2003 and predictor variables was examined. Initially, 278,445 students who were in grades 4 through 9 in the spring of 2004, who took either the ITBS or LEAP 21, and were promoted at the end of the 2003 school year were identified as initially eligible for inclusion.

Table 2*: Predicting English-Language Arts Performance: Preliminary Statewide Regression Analyses*

Predictors	Multiple correlation	Number of Students
Z-score: ELA 2003	.761	278,445
Z-scores 2003 achievement	.783	276,331
Z-scores 2003 achievement Student demographic factors	.796	275,976
Z-scores 2003 achievement School demographic factors	.755	245,878
Z-scores 2002 & 2003 achievement	.809	183,428
Z-scores 2002 & 2003 achievement Student demographic factors	.815	183,175
Z-scores 2001 – 2003 achievement	.819	125,643
Z-scores 2001 – 2003 achievement Student demographic factors	.824	125,496

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 3*: Predicting Mathematics Performance: Preliminary Statewide Regression Analyses*

Predictors	Multiple correlation	Number of Students
Z-score: Mathematics 2003	.783	279,090
Z-scores 2003 achievement	.799	276,257
Z-scores 2003 achievement Student demographic factors	.805	275,898
Z-scores 2002 & 2003 achievement	.829	183,386
Z-scores 2002 & 2003 achievement Student demographic factors	.830	183,132
Z-scores 2001 - 2003 achievement	.835	125,624
Z-scores 2001 - 2003 achievement Student demographic factors	.835	125,476

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 demographic factors. It is also clear that as the years of available achievement data increase, the contribution of demographic factors attenuates substantially. These results are nearly identical to the pilot findings from 2002-2003 with a slightly larger *n* per analysis and a higher mean level of multiple correlation. These data once again support the *potential* appropriateness of a VAA analysis based upon Louisiana's educational assessment data.

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. The various databases were linked to identify students who attended a school within a parish that participated in the LEADS pilot project and who were in grades 4-9 in the spring of 2004. Additionally, in order to contribute to these analyses, the student had be enrolled in the same school in the fall of 2003 and at the spring assessment of 2004. A substantial number of students (33,156) were identified who met these criteria.

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 three categories based upon the following criteria.

Table 4: *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.
Regularly Certified Teachers	1. Has 3 years or more teaching experience. 2. Holds an A, B, C, L1, L2, or L3 certificate.
Other/Emergency Certified Teachers	1. Teachers who are teaching on an emergency temporary authority. 2. 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. Value Added Assessment: Hierarchical Linear Model

Following the recommendations of the original pilot investigation and in order to replicate that study; the educational assessment data were analyzed using hierarchical linear models (HLM; McCulloch, & Searle, 2001; Raudenbush & Byrk, 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 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 replication was begun with no prior modeling requirements beyond those that are common to this type of analysis. In other words the models were not constrained to be similar to the models that emerged in 2002-2003. They were free to take whatever form emerged from the 2003-2004 data. It also important to note that the parish school systems participating in the LEADS project changed somewhat between 2002-2003 and 2003-2004. The 2 largest parish school systems in 2002-2003 pilot (contributing 49% of the students) were not represented in 2003-2004 data and were replaced by two different parishes. While this reduced the number of cases available for the reliability analysis, it actually strengthens the replication by providing a new sample that is more independent of the original sample.

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 prior ELA and mathematics achievement were statistically significant and were retained. In addition the random effect for social studies was statistically significant for mathematics and was retained. The models at this stage were highly similar to the prior 2002-2003 data with the addition of one to two additional random effects for each achievement domain.

In the next stage demographic co-variates were entered as a block. Limited English Proficiency did not exhibit a statistically significant random or fixed effect for ELA and was dropped from that model. For ELA the only statistically significant random effect was for special education status and this was retained. For mathematics all of the demographic variables exhibited statistically significant fixed effects and were retained. Based upon tests of significance, random effects were retained for special education status and Section 504 status. The models emerging from the 2003-2004 data are very similar to the 2002-2003 data, but slightly simpler, having fewer random effects at the student level. It is also worth noting that all of the random effects for 2003-2004 emerged in 2002-2003, it is simply the case that some of the 2002-2003 effects were not replicated in 2003-2004.

Next, the effect of a series of classroom level covariates was tested, such as class size and percentage of students who were in special education; and those that were significant were retained. Covariates were entered in the order suggested by prior *t*, with the following constraint. When the prior *t* for each of the class mean prior achievement domains (e.g., mathematics) were very similar, the prior domain that was currently being modeled entered first. This constraint was adopted to enhance conceptual clarity due to issues of shared variance. Based upon results of a significant χ^2 for heterogeneity of student level variance, 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.

The final models are presented below, followed by the results.

Table 5: *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 Gender Section 504 Status Prior year test results: *ELA*, *Mathematics*, Science, Social Studies
Classroom covariates	Students' mean prior year achievement in ELA
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 6: *Outcomes for HLM Model for English-Language Arts*

Teacher group	**Effect for Overall Achievement** In comparison to experienced certified teachers (intercept)
New University B	-8.5 (-15.6, -0.5)
New University C	-0.7 (-11.1, 9.6)
New: other universities	-13.9 (-20.1, -7.6)
Other/Emergency	-1.3 (-6.8, 4.2)

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.

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 B or the collection of other new teachers. However, the efficacy of teachers from University C was comparable to experienced teachers.

The universities represented in this year of the pilot overlap only partially with those in the prior year due to changing parish participation. The letter designations from this year match those from the prior report. Overall the results, closely parallel the results from the previous pilot year. In that year, new teachers from University B were also less effective than experienced certified teachers, however by a wider margin in the prior report. The collection of all other new teachers was similar in the degree to which they were less effective than experienced certified teachers across both pilot years.

Table 7: *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 Section 504 Status Prior Year Mathematics test result Prior year test results: *Mathematics*, *ELA*, Science, *Social Studies*
Classroom covariates	Students' mean prior year achievement: mathematics
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 8: *Outcomes for HLM Model for Mathematics Arts*

Teacher group	**Effect for Overall Achievement** In comparison to experienced certified teachers (intercept)
New University B	-8.5 (-14.2, -2.8)
New University C	3.9 (-12.3, 20.1)
New: other universities	-1.7 (-7.9, 4.5)
Other/Emergency Certified	-2.8 (-7.7, 2.1)

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.

Similar to the results for ELA, new teachers from University B were less effective than experienced certified teachers; new teachers from University C were similar to new teachers in their effectiveness. In fact, the point estimates for the effectiveness of new graduates from Universities B and C are strikingly similar to those from the previous pilot analysis.

VI. Reliability of VAA Estimates

A series of initial analyses was undertaken to examine the reliability of teacher effectiveness estimates across years. For both mathematics and ELA, reliability was estimated using those teachers who were represented in both the 2002-2003 and 2003-2004 LEADS databases. For each content area, the model that was developed in year 2 of the pilot analysis was applied to both years of the data, dropping the teacher group membership variables 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 that would tend to reduce the influence of chance factors (Sanders et al., 1997; Wright, Horn, & Sanders, 1997).

It is also worth noting that these estimates are likely to underestimate the reliability of estimates of 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 9: *Correlation between **Individual** Level Teacher Efficacy Estimates Across Years*

Teacher group	**English Language Arts** (*n*)	**Mathematics** (*n*)
All teachers	.52 (341)	.53 (359)
Teachers with 10 or more students	.55 (294)	.53 (322)
Teachers with 20 or more students	.57 (175)	.51 (229)

Generally the author would evaluate these initial *individual* level reliability estimates as very promising given that generalization across students, 12 months, and tests forms 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.

Summary

Analyses were conducted replicating the pilot work examining the feasibility of using Louisiana's educational assessment data in concert with the LEADS database and other associated databases to assess teacher preparation programs. Additionally, the reliability of an estimate of teacher effectiveness at the individual level across years, test forms, and student groups was examined. 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.

- Although there were some differences, the models were strikingly similar across years.

- The individual reliability estimates across three simultaneous dimensions of generalization were sufficient to suggest the viability of the general approach. Additional research is needed examining additional dimensions of this issue.

- As the number of years of achievement data increased, the contribution of demographic factors rapidly decreased to low levels.

- Some 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.

These analyses in concert with the prior pilot testing suggests that it may be 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. Perhaps the most striking findings are the general consistency of the models across years and the size of the preliminary reliability estimates.

A number of issues remain if this sort of modeling is to be adopted as a routine form of assessment. First, a standard model will need to be developed and employed across years. Although additional research with either the entire Louisiana database or with a representative sample will needed to accomplish this, initial two years of data suggest that this will be feasible.

A second limitation of the current analyses is that these data cannot address 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, or 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.

Additional research is needed examining, the potential contribution of cross-classified models as a means of capturing students enrollments in two classes in the same content area (i.e., two mathematics classes) and across years. Work is ongoing examining both technical and modeling issues related to this approach. These types of models are quite technically challenging to implement, but may provide the most powerful analytic tools (Rowan et al., 2002). Additional research is needed to determine what additional benefits would accrue by fitting Louisiana's data to a cross-classified model versus the models studied thus far.

In summary, the answers to the two substantive issues examined in year 2 of the pilot investigation were both supportive of continuing the VAA model development. The models were largely replicated across years and the reliability estimates are promising.

References

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.

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.

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.