Using Student Achievement Test Scores as Evidence of External Validity for Indicators of Teacher Quality: Connecticut's *Beginning Educator Support and Training* Program

Mark Wilson & PJ Hallam
University of California, Berkeley

Ray Pecheone
Stanford University

Pamela Moss
University of Michigan

Acknowledgements: We would like to thank the administrators and staff of the Connecticut State Education Departments and the two school districts who went to great lengths to help us obtain the data that were used in our analyses. We would also like to thank Hiro Yamada and Ronli Diakow for performing the analyses we report below. We would like to thank Linda Darling-Hammond for useful comments. The editor and reviewers have made many helpful suggestions, from which the paper has greatly benefited. Any errors or omissions are, of course, the responsibility of the authors. This research has been supported by a grant from the Institute for Education Sciences. Keywords: Teacher portfolio assessments, teacher standardized tests, external validity evidence

Abstract

This study examines one aspect of the validity evidence for Connecticut State Department of Education's (CSDE) performance-based teacher assessment system, the Beginning Educator Support and Training (BEST) program. Specifically, we investigate whether external validity evidence in the form of teachers' average effects on their students' achievement support the use of BEST portfolio scores as a measure of teacher quality. Using an administrative data set, the *Degrees of Reading Power* (DRP) test was used to provide evidence of student reading achievement for elementary school students in two urban school districts in Connecticut. Hierarchical Linear Modeling (HLM) findings, which take the school context into account, indicate that BEST portfolio scores did indeed distinguish among teachers who were more and less successful in enhancing their students' achievement. Specifically, a one unit change in the portfolio score corresponded to a 2.20 change in fall-to-spring DRP units, or about 46% of a year's average change, or 4 months of teaching time, for the students in this study. In an additional analysis, the relationship between the portfolio scores and alternate measures of teacher quality, ETS' Praxis series of tests, were also studied. No relationship was found between BEST portfolio scores and Praxis scores, or between Praxis scores and mean student DRP scores. These results indicate that the portfolio and Praxis assessments are measuring different constructs for these teachers. BEST portfolio scores add information that is not contained in the Praxis tests, and proved to be more powerful predictors of teachers' contributions to student achievement gains.

Using Student Achievement Test Scores as Evidence of External Validity for Indicators of Teacher Quality: Connecticut's *Beginning Educator Support and Training* Program

Licensing requirements imposed by states are designed to ensure that those who are licensed have the essential knowledge and skills to perform their work safely and "do no harm." Licensing processes are key aspects in governments' efforts to protect the public from non-qualified personnel. In the teaching profession, licensing is controlled through state governments' approval mechanisms for teacher preparation programs and teacher certification requirements. The importance of gathering evidence about classroom practice in making teacher licensure decisions was highlighted in a report from the National Research Council (Mitchell, Robinson, Plake, & Knowles, 2001). The authors of the report concluded that "paper and pencil tests provide only some of the information needed to evaluate the competencies of teacher candidates" and called for "research and development of broad based indicators of teaching competence", including "assessments of teaching performance in the classroom" (p. 172).

Acting in response to this call, this study examines one part of a validity argument for Connecticut State Department of Education's (CSDE) performance-based teacher assessment and licensure system, *Beginning Educator Support and Training* (BEST). Specifically, we investigate whether external validity evidence in the form of teachers' mean effects on their students' achievement growth supports the use of BEST portfolio scores as a measure of teacher quality. In an additional analysis, we evaluate the relationship between teachers' performance on the BEST portfolio and their performance on the Educational Testing Service's Praxis I and II tests, as well as the relationship between teachers' Praxis scores and their student learning gains.

In contrast to traditional paper-and-pencil tests, the BEST assessment provides structured evidence about beginning teachers' practice in the classroom, including their planning and teaching of a unit of instruction, student work from that unit, and commentaries on the rationale for planning and teaching decisions, as well as the teaching and learning processes and outcomes that took place during the unit. Trained raters evaluate the videotapes and artifacts of instruction in key competency areas against specific standards. These data are the basis for a decision about whether the teacher has met the standards to be granted a continuing professional license.

Validity Issues in Teacher Assessment

Performance-based evaluations such as portfolio assessments have often been advanced on the grounds of content validity (Popham, 1990), as they address some of the perceived limitations of standardized, multiple-choice tests, such as the oversimplification of teaching activities and the possibility of multiple "correct" responses for different teaching contexts (Cochran-Smith, 2003; Danielson & McGreal, 2000; Darling-Hammond, 2001b; Glass, 2002). But in order to successfully judge the usefulness of portfolio assessment in state certification systems, evidence of other forms of validity is needed as well (Herman, Aschbacher, & Winters, 1992; Kane, 2005; Moss, Schultz, & Collins, 1998; NRC, 2001). In particular, evidence of validity based on external indicators of teachers' competence is valuable, and is the focus of this study.

In teacher assessment, as in other areas in education and psychology, validity is examined following the professional guidelines in the widely accepted *Standards for Educational and*

Psychological Testing (AERA, APA, & NCME, 1999) as, "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (p. 9). In the next section we outline the validity argument this study seeks to test, and then we examine validity evidence relevant to both portfolio assessments of teaching and more traditional assessments, such as the Praxis Tests that figure in our study.

The Validity Argument

The general argument that is used to study test validity evidence through relationships with external variables is an examination of whether there is a relationship between the scores on the instrument one is investigating (in this case the BEST portfolio scores) and a measure of a variable (in this case, improvement in student performance on a test) that one believes should be correlated with the underlying variable the instrument purports to measure (in this case, teacher quality or competence). The argument takes the following form, where we are investigating evidence that an instrument, *X*, measures an underlying variable *V*, and we invoke an external variable *Y*:

(1) It is assumed known that external variable *Y* is associated with underlying variable *V*

And if

(2) We observe that instrument *X* is associated with external variable *Y*

Hence

(3) We have evidence that instrument *X* is a measure of underlying variable *V*.

Some considerations need to be borne in mind when considering this argument. For one, the argument, as it is used in the field of measurement, is based on association, not causality. In (1), this is because the nature of the relations can have may have many forms, including *V* causing *Y*, *Y* causing *V*, both of them being caused by another variable, or something more complex. In (2), this is because, for the instrument validity argument to hold, an instrument is not logically required to *cause* the external variable, it is sufficient that it be an indicator of it. Note that the nature of the evidence about the external variable (*Y*) could take two forms: (a) the external measure might itself be either established or assumed to be a measure of the underlying variable; or (b) the external measure might be a measure of another (external) variable that is not the underlying variable itself, but that one has confidence should be correlated with the underlying variable.

Thus, in the case of the BEST portfolio scores, where teacher quality is the underlying variable, an investigation of the first kind might look into the relationship of the BEST portfolio scores with scores from other instruments used to assess teacher quality in practice. Such measures are quite rare, with the Praxis III test that has been developed by Educational Testing Service (ETS; Educational Testing Service, 2008b) as perhaps the only example that is relatively widely used. However, this test was not available for use in this study.

Note that sometimes the validity argument is not quite so straightforward. Thus, had there been two instruments that were seen as measuring exactly the same underlying variable, but differ in some significant way, such as cost of implementation, etc., then a high correlation will be seen as positive external validity evidence. An example of the second kind of investigation would look into other, external, variables that one could reasonably assume should be associated with teacher quality. One can classify these external variables into three possible types:

(i) other instruments intended to measure other teacher qualities, that one might expect to be correlated with the original variable, such as, say, supervisor or principal ratings;

(ii) expected outcome variables that one might expect to vary by teacher quality such as (a) teacher effects on their students' achievement (which might be operationalized using a gain-score, or by analyzing the residuals on a post-test after allowing for a pre-test, and other possible predictors), or (b) teaching process variables, such as whether the teacher uses teaching practices believed to be associated with improved student learning (which might be operationalized by observing the teacher and noting the frequency and/or quality of such practices); or

(iii) predecessor (or "exogenous " or "upstream") variables such as teacher background variables that one might expect to be associated with higher teacher quality, such as greater command of the subject matter (which might be operationalized by noting the years of study of that subject at the college level), or measures of teacher background knowledge such as other ETS Praxis tests (e.g., PRAXIS II, which is designed to measure pedagogical content knowledge, or even PRAXIS I, which is designed to measure core skills in reading, writing and mathematics).

As with the first kind of investigation, the validity evidence that is gathered is correlational in its nature: It may be that there is a complex causal structure involved, but teasing out the causal links is not seen as a required step.

Not all of these forms of evidence are seen in an equal light. For example, one might consider teacher background variables such as undergraduate qualifications etc., to be too distal from the underlying variable, teacher quality, to be of much value. Or, one might be concerned that teaching process measures are not solid evidence because one cannot be sure that the teacher employs them consistently outside of the observational situation. Facing these arguments, some turn to outcome variables of teacher quality, the effects that they have on their students, as the ultimate form of evidence. It is in this spirit of rigorous examination of validity evidence that we have turned to the investigation of correlations with student test scores as one important aspect of validity evidence for the BEST portfolio scores.

Turning back to the form of the validity argument given above, in the case at hand, where the instrument is the score on the teacher's BEST portfolio, the underlying variable is teacher quality, and the external variable is the effect that the teacher has on student outcomes, the specific argument is as follows:

(4) It is *assumed known* that the effect that a teacher has on student outcomes is associated with teacher quality

And if

(5) We *observe* that the BEST portfolio score is associated with the effect that a teacher has on student outcomes

Hence

(6) We *have evidence* that the BEST portfolio is a measure of teacher quality.

Note that there are threats to this logic that we will need to keep in mind and investigate in this study. For example, regarding (4) one might speculate that variables other than teacher quality

could affect the particular measures of student outcomes that are being used. For example, there may be teacher sorting with respect to the initial values of student achievement—i.e., teachers with indicators of higher quality tend to be assigned to students with higher test scores, and hence, their students would be expected to have higher gains--(Clotfelter, Ladd & Vigdor, 2006). Hence, this threat to the logic will need to be considered.

The primary research questions of this study are the following: (a) To what extent do the BEST portfolio scores allow one to distinguish among teachers who are more and less successful in enhancing their students' achievement? In addition, we explore whether other variables at the student level are influential in this relationship. This question corresponds to subtype (ii) of the second kind of investigation described above. And (b) To what extent do the portfolio scores allow one to distinguish among teachers who have higher and lower scores on a standardized test of teacher knowledge? This question corresponds to subtype (iii) of the second kind of investigation described above.

Research on the Validity of Teacher Assessments

In general, statistically significant and important findings are often difficult to achieve in research on the relationship between teacher characteristics and student achievement. Milanowksi (2004) did find such relationships between teachers' scores on standards-based observation tools and their students' learning, but he points out that "It is important to recognize that very high correlations between teacher evaluation scores and student achievement measures are unlikely to be found for reasons including error in measuring teacher performance, error in measuring student performance, lack of alignment between the curriculum taught by teachers and the student tests, and the role of student motivation and related characteristics in producing student learning" (p. 50).

Glass (2002) concluded that traditional paper-and-pencil tests of ability and achievement have generally failed to predict teaching effectiveness in terms of student achievement. However, some studies have found that teacher licensing test scores, along with variables like certification in the field taught, are significant predictors of student learning gains (see for example, Clotfelter, Ladd, & Vigdor, 2007; Ferguson, 1991).

Since PRAXIS tests (in particular PRAXIS I and II) play a role in this study as alternate indicators of teacher quality, we review briefly the results of validity research regarding these tests. Much of this research deals with the content validity of the tests; that is, whether there is evidence that they measure the knowledge and skills intended. For example, the Educational Testing Service (ETS, 1999) carried out a Praxis validation effort linking the knowledge and skills measured by the tests to the jobs of entry-level teachers. A job analysis was conducted to define the content-related background knowledge and skills that all newly licensed elementary school teachers (grades K-6) should possess in order to perform their job. Various strategies are used to ensure that the test items measure knowledge and skills important for entry-level teachers in the state, including reviews by committees of subject-matter experts, multiple expert reviews, and other verification procedures from each state that adopts a Praxis assessment (Rosenfeld & Kocher, 1998).

Powers (1992) analyzed surveys on classroom performance criteria from two multi-state samples of educators. These survey respondents provided ratings of the importance of

preliminary versions of the criteria being developed for the Praxis series. Powers found considerable agreement about the importance of the criteria across several classifications of educators according to ethnicity, instructional level, years of teaching experience, subject area, and orientation to teaching. A wide-ranging review conducted under the auspices of the National Research Council concluded that the evidence collected on the Praxis series exhibited a reasonable level of psychometric validity, "With a few exceptions, the Praxis I and Praxis II tests reviewed meet the criteria for technical quality articulated in the committee's framework" (NRC, 2001, p.87). However, the NRC review did not find any positive evidence of the relationship between student achievement and either the Praxis I or Praxis II tests.

Crehan & Mikitovics (2002) examined ETS' Pre-Professional Skills Test (PPST), a test of basic skills in math and reading similar to Praxis I. They found that the correlations between the PPST scores and student teaching grades were negligible and not statistically significant. Correlation and hierarchical regression findings did not suggest a predictive relationship between PPST scores and student-teaching ratings, and indicated only a weak predictive relationship between PPST scores and undergraduate GPAs.

A study of ETS teacher tests by Selke, Mehigan, Fiene & Victor (2004) compared standardized basic skills (reading, writing, mathematics, and grammar) and content area test scores with performance items on an INTASC standards-based rubric, scored by beginning teachers' supervisors. The authors concluded that there was no correlation between any aspect of the content area tests and classroom performance of first-year teachers as assessed by immediate supervisors.

A more recent study did find a positive relationship between some Praxis tests and student achievement (Goldhaber, 2007). Teachers who met North Carolina's Praxis II requirements were more effective in math, and marginally more effective in reading. Further, the higher teachers scored on the Praxis CIA exam, the higher student achievement scores were in literacy and math. In general, these patterns were found for both black and white teachers and for the various subgroups of students. To address the issue of nonrandom matching of teachers and students, Goldhaber used models that included school and student fixed effects. Teacher effects were identified based on variation in teacher qualifications within schools across classrooms and across student over time. Results show that the nonrandom sorting of teachers did have an impact on the estimated relationship between teacher test performance, however, the findings still suggested that performance on the PRAXIS tests provided a weak signal of teacher effectiveness.

Validity studies of portfolio assessment of teachers in certification processes have become more frequent over the past few years. For example, researchers have examined the constructs and impact of the National Board for Professional Teaching Standards (NBPTS), a certification process designed to identify accomplished teaching. Most of these studies have found that Board-certified teachers are associated with larger value-added achievement gains for their students than teachers who attempted the certification process and failed or other teachers of similar experience working with similar students (Bond, Smith, Baker, & Hattie, 2000; Cantrell, Fullerton, Kane, and Staiger, 2007; Cavaluzzo, 2004; Clotfelter, Ladd, and Vigdor, 2007; Goldhaber & Anthony, 2005; Goldhaber , 2007; National Research Council, 2008; Smith, Gordon, Colby, and Wang, 2005; Vandevoort, Amrein-Beardsley, & Berliner, 2004). Two studies reported mixed results, with students of Board-certified teachers showing stronger gains

than other students in some instances but not others (Harris and Sass, 2007; Sanders, Ashton, and Wright, 2005).

Despite the range of studies that have been conducted, the issue of the assessment's validity has been debated (Bond, 2001; Cunningham & Stone, 2005; Podgursky, 2001), highlighting a need for further exploration of the relationship between student achievement and teacher portfolio assessment. In reviewing 11 such studies, a recently completed NRC study concluded: "We see a relationship between board certification and student achievement, although the relationship is not strong and is not consistent across contexts (NRC, 2008).

Studies of another assessment -- the Performance Assessment for California Teachers (PACT) that is designed to identify competent teacher credential candidate – have provided several strands of evidence of related to validity and scoring reliability (Pecheone and Chung, 2007; Pecheone, Pigg, & Chung, 2005), but to date have not yet completed research on the relationship of PACT teacher scores with external criterion validity such as student achievement.

Measures Used in this Study

Connecticut's *Beginning Educator Support and Training* (BEST) Assessments

At the time the study was conducted, there were three levels of teacher licensure in Connecticut. To receive an initial license, a teacher had to pass appropriate PRAXIS tests (as well as fulfill other teacher education requirements). To receive a continuing professional license (at year three), the teacher had to engage in the BEST induction program, and pass the BEST portfolio assessment. A third level required additional professional requirements. The BEST program was a two to three year comprehensive program of support and assessment. The support component consisted of mentors or support teams from the teachers' own school or district, who successfully participated in state sponsored support training.

The portfolio assessment component required teachers in their second year of teaching to submit a content-specific teaching portfolio. In this study, the content area is "Elementary Education" (EE), and the participants were 3rd through 6th grade multiple-subject teachers (CSDE, 2006). EE portfolios require documentation of five to eight hours of instruction on one literacy unit and one mathematics unit for one class of students. Documentation includes teacher lesson plans, videotaped segments of teaching, student work, and reflective commentaries on the teaching and learning that took place during the unit. Due to constraints on the acquisition of appropriate student data, only the literacy scores for the portfolios were analyzed[1].

In the BEST program, beginning teachers were required to demonstrate, through the portfolio assessment, acceptable levels of essential teaching competencies related to the four domains of teaching (a) instructional design, (b) instructional implementation, (c) assessment of student work, and (d) teacher reflection on their practices and student outcomes, including indicators of student learning. Beginning teachers who did not successfully complete the portfolio assessment in year two were required to submit a portfolio in their third year of

[1] Reading comprehension (via the DRP) was the only subject that school districts consistently assessed for all students in both the fall and spring. Thus, collecting appropriate data on student achievement in mathematics and writing was not possible.

teaching. For the purposes of this study, each teacher's first BEST score was used in data analyses.

As implemented at the time of our study, the portfolios were evaluated by experienced teachers who have received at least five days of training and passed a calibration test based on pre-evaluated benchmark portfolios. Each portfolio was evaluated independently by two assessors, and where significant differences were found, a third reader was called in to reconcile the scores. Assessors first took notes on the portfolio based upon a series of guiding questions (GQ's) also provided to the beginning teacher. The questions were organized into four categories: instructional design, instructional implementation, assessment of learning, analyzing teaching and learning. Then assessors decided on one of four performance levels based upon an integrative scoring rubric that characterizes the performance levels and describes the associated consequences. Assessors review their notes and cite evidence for each guiding question to arrive at a score as well as complete a "feedback rubric" which contains performance level descriptions for each guiding question and is used to give more specific feedback for the beginning teacher.

All portfolio notes and scores were audited by an assessor trainer who provided additional training if readers seemed to be drifting from the benchmarks. Independent re-evaluations were conducted for all failing portfolios, as well as for a sample of just-passing portfolios (2 on a 4 point scale), for any portfolios where readers could not agree on the score, and for any portfolios where the trainer did not feel the notes justified the score given. Judges were expected to score approximately 2 portfolios per day. Reliability information was routinely maintained based upon the initial scoring by two assessors and the independent audited rescores. Studies indicated that the inter-rater reliability coefficients for the portfolios were at acceptable levels (r = .72 to .76) (Pecheone & Stansbury, 1996; Youngs, 2002).

Policy capturing techniques were used to establish passing standards. An independent committee of teachers reviewed actual portfolios to develop the descriptions of the performance levels and selected benchmarks and then a second committee independently confirmed pass/fail decisions on pre-evaluated portfolios blind to their pass/fail status. Before a portfolio assessment for a particular subject area went on line, the state conducted a special reliability study where a sample of portfolios are scored by multiple pairs of readers and commissioned an independent audit of the development process and alignment among standards, portfolio handbooks, scoring materials, and training procedures.

Traditional studies of external validity evidence had not been made on BEST portfolios at the time of this study; however, other studies of the BEST program provided insights regarding its use. A study of beginning science teachers in Connecticut by Lomask, Seroussi, and Budzinsky (1997) reported that teacher participants in this pilot science performance assessment indicated that most teachers who participated in portfolio development and the program's support seminars found it to be a good opportunity for reflection and professional growth. Wilson, Darling-Hammond & Berry (2001) suggested in their study on the BEST system that Connecticut students' increasing scores on the Connecticut Academic Performance tests and high scores on the 1998 NAEP reading test - despite an increase of the state's poverty index by nearly 50% - could be attributed in part to "the harvest of this work" (p. 28).

The CSDE's *Praxis* Tests

CSDE provided data on both PRAXIS I and II tests for use in this study. CSDE requires two examinations: (a) Praxis I: Academic Skills Assessments, which are designed to measure basic proficiency in reading, mathematics, and writing, and (b) Praxis II: Subject Assessments, which are designed to measure content area knowledge. All individuals seeking (a) formal admission to a teacher education program or (b) licensure, must take and pass the Praxis I: *Pre-Professional Skills Tests in Reading, Writing, and Mathematics*, or meet the requirements of one of the State Board-approved SAT waiver options. The Praxis I test consists of four sections: (a) math, (b) reading, (c) writing – analysis, and (d) writing – essay. The first three sections have a multiple choice format, and the fourth is an on-demand essay written to a prompt.

For elementary teachers, the Praxis II tests that were required at the time of this study were the *Curriculum, Instruction & Assessment* (CIA) and *Content Area Exercises* (CAE). These assessments were designed to measure general pedagogical knowledge at the K-6 level. The tests used multiple-choice items and featured a case study approach with constructed responses. Test-takers who fail Praxis I or II are allowed to re-test at a later date. In this study, teachers' first Praxis I and Praxis II scores were used.

Praxis multiple choice questions are machine-scored. Scoring reliability was ensured through ETS' professional scoring practices (ETS 2008a). Raters score the essay and constructed response portions of Praxis using a holistic method of evaluating the overall quality of thinking and writing against Praxis standards. Raters must have at least a Bachelor's degree in the field that they score. ETS trains raters through their interactive tutorial website, and they must pass rater consistency tests.

The *Degrees of Reading Power* (DRP) test

The school districts that provided the data used in this study routinely administered the *Degrees of Reading Power* test (Touchstone Applied Science Associates, 2006) in the fall and spring of every school year These student scores provided pre and post testing data for this study. The DRP is a standardized reading achievement test that uses a modified cloze technique (filling in missing words from a phrase) to assess reading comprehension (Touchstone Applied Science Associates, 2006). Findings from a study of the psychometric properties of the DRP indicated that it has high level of reliability (test-retest = .95), construct validity, and criterion-related validity (Koslin, Zeno, & Koslin, 1987). An advantage of the DRP for researchers is that interval scale scores are available for all forms and levels of the test. According to the publisher of the DRP test, a year's growth usually falls in the range of 8-10 units (Touchstone Applied Science Associates, 2005).

Data and Methods

The Data Set

The data set constructed for this study combined student-level administrative data from two school districts with state administrative data about the teachers linked to those students. The data set included all the teachers (and their associated data) that fit our profile and whose data were administratively available,

Two urban Connecticut districts were selected on the basis of (a) their willingness to

allow data to be used for this project, and (b) their routine practice of including a spring administration of the state's DRP test in addition to the state's fall administration which allowed us to consider student achievement change as a variable. A superior design would involve randomization among teachers and students, but this was beyond the scope of this study, which is based on administratively-available data. Information about teachers and their students was collected under approved guidelines of the Institutional Review Boards at the University of Michigan and the University of California, Berkeley, and following the guidelines for the CSDE as well as the two school districts. All appropriate efforts were undertaken to guarantee anonymity of the teachers and students whose data were used in the analyses reported here.

CSDE provided data about teachers from the two districts from the past four school years for 104 3rd, 4th, 5th, and 6th grade teachers who completed BEST portfolios. These datasets include the following information about teachers: (a) overall portfolio scores, (b) their scores on Praxis I and II tests, and (c) demographic data. Only teachers who had spring and fall data for the students in the class and a completed BEST portfolio were included in the data set.

Table 1 provides descriptive data for the teachers in this study. The teachers in this study are mostly female (84%) and white (72%), as is typical for teachers in the U.S. (National Center for Education Statistics, 2004). The plurality of teachers taught 4th grade, 36%, but they are fairly evenly spread across the four grades.

Table 1
Teacher Gender, Ethnicity, District and Grade Levels[a]

Gender			Ethnicity			District			Grade		
	n	%		n	%		n	%		n	%
male	16	15	African Am.	10	9	1	61	55	3	21	19
female	88	80	Euro. Am.	73	66	2	43	40	4	44	40
NR[b]	6	6	Hispanic	19	17	3	6	5	5	22	20
	11				7		11				
total	0		NR	8		total	0		6	23	21
				11							
			total	0					total	110	

[a] Note: percentages may not always add to 100 because of rounding.
[b] Note: NR indicates "no response."

The student data were provided by the two school districts. The results in Tables 2 and 3 indicate that almost half of the students were African American, and over one-third were Hispanic. Almost all of the students qualified for free or reduced lunch, indicating that the students in this study are from high poverty families. The percentage of students that qualified for special education was 11%, and 13% qualified for English Language Learners services. Several of the categories have fairly large proportions of students with missing data for these categories: consideration of this will be included in the analyses.

Table 2
Student Ethnicity and Lunch Status[a]

Ethnicity			Lunch		
	n	%		n	%
Native Am.	78	4	Free	1625	78
Asian Am.	21	1	Reduced	267	13
African Am.	980	47	Full	175	8
European Am.	204	10	NR	20	1
Hispanic	750	36	Total	2087	
Other	52	3			
NR[b]	2	0			
total	2087				

[a] Note: percentages may not always add to 100 because of rounding.
[b] Note: NR indicates "no response."

Table 3
Student Gender, Special Education Status, and English Language Learners[a]

Gender			Special Ed			ELL		
	n	%		n	%		n	%
Male	1071	51	yes	244	12	yes	260	12
Female	1014	49	no	1112	53	no	1463	70
NR[b]	2	0	NR	731	35	NR	364	17
total	2087		total	2087		total	2087	

[a] Note: percentages may not always add to 100 because of rounding.
[b] Note: NR indicates "no response."

Covariates

Absent a randomized design for data collection, we sought to control for as many potentially confounding variables as possible by including them as covariates in the analysis. At the student level, students' socio-economic status is consistently found to be a factor in student achievement. In this analysis, we used Lunch Status (free/reduced/full) as a proxy for socio-economic status. Other aspects of student background that have been found to be associated with student achievement are gender, English-language learner status and special education status (Darling-Hammond, 2000; Ehrenberg & Brewer, 1995; Wenglinsky, 2003). All three are available in the data set, and so were included in the analysis.

Where there were instances of missing data in the administrative data set, we coded them as "missing," in the cases where there was very little missing data (e.g. 1 or 2 cases). In cases where there were greater amounts of missing data, we included a separate "missing data" variable for each such covariate in order to evaluate whether the lack of data might be associated with factors that were potentially influential,

Correlational Analyses

Three correlation analyses were completed using traditional two-tailed Pearson

calculations. The first analysis examined the relationships among BEST portfolio scores, Praxis I scores, Praxis II scores and student gain scores. The second correlated BEST portfolio scores with Praxis I and II scores. The third analysis used a partial correlation, holding the pre-test scores constant to correlate student post-test scores with (a) portfolio scores and (b) Praxis II scores.

Hierarchical Linear Modeling (HLM)

We used hierarchical linear modeling to examine the impact of teacher characteristics on student achievement because it can help sort out the magnitude of impacts at different levels of the education system from which influences on student learning emerge (Bryk & Raudenbush, 1988).

We use pre-test scores as a covariate for the spring achievement scores. Although the idea of gain scores is intuitively appealing and a more straightforward method to explain to many audiences, the use of the pre-test as a control is preferred for several reasons. Bereiter (1963) points out that there are three problems regarding the gain score approach. First, the reliability of gain scores is inversely related to the pretest-posttest correlation; that is, the higher the reliability of gain scores, the lower the pretest-posttest correlation, and vice versa. This can lead to difficulties in interpreting difference scores as change. Second, gain scores may not be on the same scale for persons at different levels of initial scores. For example, a gain score of 3 for a person with a high initial score, for example, 75 on the DRP scale, may have a different meaning from the same amount of gain for a person with a lower initial score, such as 25. Third, gain scores are spuriously negatively correlated with pretest scores since the same errors of measurement, with an opposite sign, are present in both scores being correlated. This negative correlation, as Lord (1963) indicates, exists due to the effect of regression toward the mean, even if no real change has occurred. In their review of the estimation procedures of difference scores such as raw gain or residual gain, Cronbach and Furby (1970) came to the conclusion that it is better to avoid estimating gain scores and to take alternative methods such as a covariance approach that takes various indices of initial status into account.

Accordingly, a two-level linear modeling analysis was conducted to investigate teacher effects on student achievement. These analyses were conducted in terms of the post-test scores, using the pre-test scores as a covariate. These analyses were conducted with the following additional covariates at the student level: student initial status, ethnicity, gender, free lunch status, special education status, and ELL status. For the teacher level the following variables were used: Teachers' BEST portfolio scores and Praxis scores. Teacher-level covariates include teacher demographic data and type of mentoring program, and prestige of undergraduate institution.

A random intercept HLM model was used to examine whether there are statistically significant and important associations between teacher performance and classroom student achievement, using STATA (2005). Empirical Bayes estimates increase the reliability by weighting the more reliable data more heavily. This model is preferable to ordinary least squares estimates of residuals especially for this study because teachers' classes had varying sample sizes. By using a random intercept model, each teacher's class of students can have its own intercept, providing information about the percentage of variation in outcomes at both levels (i.e., student and teacher levels). Note that the DRP results were not standardized before analysis:

This approach was chosen so that the results could be presented in terms of DRP Units, which have useful interpretability.

As there was missing data shown in Table 3 at the student level, we included a missing data category as well for each variable with significant missing data (i.e. more than one or two cases). Including them as a separate code allows us to gauge whether their presence affects the basic findings.

A random intercept HLM model was used to examine whether there are statistically significant and important associations between teacher performance and classroom student achievement, using STATA (2005). Empirical Bayes estimates increase the reliability by weighting the more reliable data more heavily. This model is preferable to ordinary least squares estimates of residuals especially for this study because teachers' classes had varying sample sizes. By using a random intercept model, each teacher's class of students has its own intercept, providing information about the percentage of variation in outcomes at both levels (i.e., student and teacher levels).

We examined whether adjustments might be needed for the effect of teacher sorting, which would be evidenced by a positive correlation between initial student achievement and teacher scores. The approach described by Goldhaber and Anthony (2004) and Clotfelter et al (2006) used fixed effects to control for this effect. In this data set, however, there is no significant correlation between initial student achievement and teacher scores, -0.102 (p= 0.3008), revealing that the phenomenon observed by these researchers is not indicated for this data set—hence, we use the more straightforward HLM approach.

Prior to conducting the HLM analysis, BEST data were analyzed for internal reliability. Table 4 shows that there was high reliability for each domain created via aggregation.

Table 4
Teacher Level Composite Variables: Reliability

Variables	Cronbach's Alpha
Average Instructional Planning	.86
Average Instructional Implementation	.82
Average Student Assessment	.84
Average Reflect on Practices	.88
Teacher preparation program	.82
Teacher education faculty	.84
Cooperating teacher	.81
Support for preparing portfolio	.78
Opportunity to show via portfolio	.94

Results

Student Achievement

Overall, the data indicate that student achievement in reading comprehension varied across a wide range, and that the majority of students in this data set increased their reading comprehension to a modest extent. Students' posttest scores on the DRP ranged from a low of 15 to a high of 95. The students' mean posttest score was 44, which is in the expected range of 3rd grade scores. A large majority (71%) of the mean posttest scores fell between 30 and 60. According to TASA's (2006) DRP Scale of Text Difficulty, these scores indicate the majority of students were in the "Primary School Textbook" range (3rd to 4th grades) represented by books such as *Green Eggs and Ham* (Level 31) to *Charlotte's Web* (level 50). The range of DRP scores also dips below this range. But 22 of the student posttest scores ranged from 80 to 95, which aligns with the "High School Textbook" levels and above. Thus, the chosen outcome variable, DRP score, represents a variable that has educationally significant variability, which is important in valuing the analytic results.

Correlations

Findings from the correlation analysis of BEST portfolio scores and Praxis scores are presented in Tables 5 and 6. The Praxis tests correlated with mean gain scores at levels ranging from -.10 to .01. Teachers' portfolio scores correlated with mean student gain scores at a slightly higher level (.17), but this was not statistically significant. (See table 5.) Results for partial correlations, controlling for fall DRP scores, were also calculated with similar results. Table 6 shows that there was no relationship between BEST portfolio scores and the three standardized tests of teacher knowledge offered in the Praxis series, indicating that the tests are measuring different constructs.

Table 5
Correlations of Teacher Assessments and Mean Student Gain Scores

Assessment	Correlations	Approx. Std err	Sample size
Portfolio	.17	0.10	110
Praxis I	-0.04	0.13	57
Praxis II (CIA)	-.010	0.11	101
Praxis II (CAE)	.01	0.10	98

Table 6
Correlations of Teachers' Portfolio Scores and Praxis Scores

	Correlation	Approx. std err	Sample size
Praxis I Mean	-.15	0.13	57
Praxis II CIA	-.11	0.11	101
Praxis II CAE	.01	0.10	98

HLM Findings

The outcome variable in our HLM analyses is DRP post-test score, with DRP pretest score always included as a level 1 (student) covariate. Table 7 indicates that seven of the Level 1 covariates were statistically significant. The most highly significant covariate was DRP pretest scores ($z = 26.56$; $p < 0.001$), which would be expected. The others were (a) Special Education ($z = -5.30$; $p < 0.001$), and (b) Special Education Missing ($z = -4.40$; $p < 0.001$), Free and Reduced Lunch Status ($z = -3.44$; $p < 0.01$), Grade ($z = 3.31$; $p < 0.01$), English Language Learner status ($z = -3.02$; $p < 0.01$), and English Language Learner Missing ($z = -2.82$; $p < .01$). As speculated above, missing data status was indeed statistically significant for some of the student variables: Special Education and English Language Learner. It is important to our main interest to control for these effects, but, unfortunately, it is difficult to interpret the effects themselves in this administrative data set—one could speculate as to why they are statistically significant, but the reasons for the missing status are not available to us. The European American ethnicity variable also comes quite close to significance at the standard α=0.05 level, and did indeed reach that in some of the preliminary analyses that were done before this final analysis.

We use as an effect size indicator the proportion of variance accounted for (R^2) derived from comparing the model with Level 1 and Level 2 covariates with a null model (i.e., one with no covariates). The amount of variance accounted for at the *student* level (R^2_W or the variance within), 0.32, indicates about a third of the variance at Level 1 is explained by Level 1's student covariates. This gives a comparison for the amount of variance explained by teacher variance. The intra-class correlation coefficient (ICC) indicates what percent of <u>total</u> variance was due to Level 2 (teacher) variance. High ICC values would indicate that Level 2 covariates contributed a great deal to the variance between students' pre and post test scores. The ICC for this model was 0.18, which indicates that the teacher level did contribute to the variance, a little more than half that explained by the student-level variables, although there is still a considerable amount of the variance not explained by the teacher level. In contrast, the amount of variance accounted for at the <u>teacher</u> level (R^2_B or the variance between), 0.80, indicates that a great deal of the variance at Level 2 is explained by the teacher level covariates.

Among the nine Level 2 covariates, the only significant coefficient is for the BEST portfolio scores (coefficient = 1.54; $p = 0.03$). (The other eight covariates, which prove to have little influence on student achievement, include Praxis scores, prestige of the teachers' preservice institution, race and gender.) The size of the effect of the portfolio scores on student achievement is substantial. Specifically, one unit change in the portfolio score corresponds to a 2.20 change in DRP units, or about 46% [= 2.20/4.8] of a year's

average change for these students (i.e., about 4 months of teaching time).

This finding, which is substantially different from the finding of the simpler correlational analyses reported above, and arguably a better representation of the results, supports claims that HLM analyses are superior to traditional forms of analysis affects on student achievement (Wenglinsky, 2002). The multivariate analysis, with its greater statistical controls, and the ability of HLM to account for school and teacher level effects, better represents the independent effects of this measure of teacher quality.

Table 7
HLM Results with Post DRP Scores as Outcome

	Covariates	Coef.	SE	z	P>|z|
Level 1: Student	Pre DRP	0.62	0.02	35.08	0.00
	Grade	-0.17	0.44	-0.40	0.69
	ELL	-1.83	0.69	-2.67	0.01
	Female	0.02	0.38	0.05	0.96
	African Am.	-0.53	1.19	-0.44	0.66
	European Am.	2.40	1.24	1.95	0.05
	Hispanic	0.52	1.20	0.44	0.66
	Lunch status	-1.66	0.53	-3.12	0.00
Level 2: Teacher	Portfolio score overall	1.54	0.71	2.17	0.03
	Female	0.61	1.41	0.43	0.67
	European American	-1.34	1.22	-1.10	0.27
	Independent mentoring	-1.81	1.14	-1.58	0.11
	Urban 2 districts	0.65	1.13	0.57	0.57
	Praxis II: CIA	-0.04	0.05	-0.81	0.42
	Praxis II: CAE	-0.01	0.05	-0.27	0.79
	Prestige of teacher pre-service institution	1.16	1.58	0.73	0.46
	Pass Praxis on 1st try	2.16	1.57	1.37	0.17

Note. The number of students and teachers available in these analyses was 1466 and 85, respectively

Conclusions

Licensure processes serve the public's interest by providing a framework for selecting qualified, competent practitioners (Kane, 2005). If they are useful in pursuing this goal, certification tests should differentiate those who can practice successfully from those who cannot. Findings on this study's first study question, "To what extent do the BEST portfolio scores allow one to distinguish among teachers who are more and less successful in enhancing their students' achievement?" indicate that this assessment does indeed allow us to distinguish among teachers who were more and less successful in enhancing their students' achievement. HLM findings revealed that one unit change in the portfolio (scored on a 4 point scale) corresponded to a 2.20 change in DRP units, or about 46% of a year's average change for these students (i.e., about 4 months of teaching time). The ICC value of 0.18 indicates that portfolio performance was a reasonably large contributor to the total variance, but that there is still considerable variance unaccounted for.

Findings on our second question, "To what extent do the portfolio scores allow us to distinguish among teachers who have higher and lower scores on a standardized test of teacher knowledge?" showed that there was negligible correlation between the BEST portfolio scores and the PRAXIS scores. Thus, whatever is the aspect of the BEST scores that is associated with the improvement in student scores, it is not shared with either of the PRAXIS tests. Together, these results can be summarized as indicating that there is evidence for the external validity of the BEST portfolio, and this association is not due to shared variance with the PRAXIS I or II standardized tests. This finding is bolstered by the finding reported in Table 4 that none of the PRAXIS tests correlate with DRP change or with the BEST portfolio. We see this as an important result for both practitioners and researchers in the area of teacher assessment.

There are several limitations to this study that need to be borne in mind when interpreting the results. The study is based on a secondary data analysis. The data were originally collected for other purposes, and then linked for the purposes of this research. Hence, there was no opportunity to apply randomization of any kind to strengthen the design. Nevertheless, given the strictures of using data from a state-run licensure program, the project did undertake stringent means to ensure data integrity, particularly the integrity of the links between student and teacher data. Second, missing data may not have been missing at random, as required by the HLM approach. As Braun (2005) noted, incomplete data from districts may contribute to possible sources of bias. However, we did include missing data as a category in the analyses, and this helped sensitize the results to this issue.

This study is one of the first to examine the external validity of licensure assessments as predictors of later teacher success in promoting student learning. It is our hope that this kind of research will become more commonplace, and that larger-scale studies with more extensive data bases will be conducted to enhance our understanding of how assessments can capture the essential knowledge and skills needed for successful teaching.

References

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Ballou, D., & Podgursky, M. (1999). Reforming teacher preparation and licensing: What is the evidence? Teachers College Record.

Bond, L. (2001). On "Defrocking the National Board": A Reply to Podgursky. *Education Next 1*(3), pp. 4-5.

Bond, L., Smith, T., Baker, W., & Hattie, J. (2000). *The certification system of the National Board for Professional Teaching Standards: A construct and consequential validity study*. Greensboro, NC: Center for Educational Research and Evaluation at the University of North Carolina at Greensboro.

Braun, H. I. (2005). Value-added modeling: What does due diligence require? In R. Lissitz (Ed.), *Value added models in education: Theory and applications* (pp. 40-79). Maple Grove, MN: JAM Press.

Bryk, A. S., & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education, 97*, 65-108.

Cavaluzzo, L. (2004). *Is National Board Certification an effective signal of teacher quality?* (National Science Foundation No. REC-0107014). Alexandria, VA: The CNA Corporation.

Clotfelter, C., Ladd, H. Vigdor, J. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources, 41* (4) 778-820.

Cochran-Smith, M. (2003). The unforgiving complexity of teaching: Avoiding simplicity in the age of accountability. *Journal of Teacher Education, 54*, 3-5.

Crehan, K., & Mikitovics, A. (2002). Pre-Professional Skills Test Scores as college of education admission criteria. *Journal of Educational Research, 95*(4), 215-223.

Cunningham, G. C., & Stone, J. E. (2005). Value-added assessment of teacher quality as an alternative to the National Board for Professional Teaching Standards: What recent studies say. In R. Lissitz (Ed.), *Value Added Models in Education: Theory and Applications* (pp. 320). Maple Grove, MN: JAM Press.

Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice* (No. ED455). Princeton, NJ: Educational Testing Service.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1).

Darling-Hammond, L. (2001a). *The research and rhetoric on teacher certification: A response to "Teacher Certification Reconsidered"*. New York, N.Y.: National Commission on Teaching & America's Future.

Darling-Hammond, L. (2001b). Teacher testing and the improvement of practice. *Teaching Education, 12*(1), 11-34.

Educational Testing Service. (1999). *Validity for licensing tests*. Princeton, NJ: Educational Testing Service.

Educational Testing Service. (2008a) Overview of Praxis I and Praxis II Tests. Accessed on September 9, 2008 from http://www.ets.org/portal/site/ets/menuitem.1488512ecfd5b8849a77b13bc3921509/?vgnextoid=2684c05a4d1ae010VgnVCM10000022f95190RCRD&vgnextchannel=b032c05a4d1ae010VgnVCM10000022f95190RCRD

Educational Testing Service. (2008b). *Understanding your Praxis scores* (The Praxis Series ed.). Princeton, NJ: Educational Testing Service. Retrieved June 20, 2009 from http://www.ets.org/Media/Tests/PRAXIS/pdf/uyps_web.pdf.

Ehrenberg, R., & Brewer, D. (1995). Did teachers' verbal ability and race matter in the 1960s? Coleman revisited. *Economics of Education Review, 14*(1), 1-23.

Glass, G. (2002). Teacher characteristics. In A. Molnar (Ed.), *School Reform Proposals: The Research Evidence* (pp. 1 - 6). Tempe, AZ: Arizona State University.

Goldhaber, D. (2007). Everyone's doing it, but what does teacher testing tell us about teacher effectiveness? *Journal of Human Resources, 42*(4), 765-794.

Goldhaber, D., & Anthony, E. (2004). *Can teacher quality be effectively assessed?* Seattle, WA: University of Washington and the Urban Institute.

Goldhaber, D., & Brewer, D. E. (1997). Evaluating the effect of teacher degree level on educational performance. In W. Fowler (Ed.), *Developments in school finance*. Washington, DC: US Department of Education.

Goldhaber, D. D., & Brewer, D. J. (2000). Does teacher certification matter? High school teacher certification status and student achievement. *Educational Evaluation and Policy Analysis, 22*(2), 129-145.

Hawk, P., Coble, C. R., & Swanson, M. (1985). Certification: It does matter. *Journal of Teacher Education, 36*(3), 13-15.

Herman, J., Aschbacher, P., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: ASCD Publications.

Kane, M. T. (2005). The role of licensure tests. *The Bar Examiner, 74*(1), 27-38.

Koslin, B. L., Zeno, S., & Koslin, S. (1987). *The DRP: An effective measure in reading*. Brewster, NY: TASA DRP Services.

Laczko-Kerr, I., & Berliner, D. C. (2002). The effectiveness of "Teach for America" and other under-certified teachers on student academic achievement: A case of harmful public policy. *Education Policy Analysis Archives, 10*(27).

Ladson-Billings, G., & Darling-Hammond, L. (2000). The validity of National Board for Professional Teaching Standards (NBPTS)/Interstate New Teacher Assessment and Support Consortium (INTASC) assessments for effective urban teachers: Findings and implications for assessments. On *Educational Resources Information Center (U.S.)*. College Park, MD Washington, DC.

Lomask, M., Seroussi, M., & Budzinsky, F. (1997). *The validity of portfolio-based assessment of science teachers.* Paper presented at the National Association of Research in Science Teaching, Chicago, IL.

Lustick, D., & Sykes, G. (2006). National Board Certification as professional development: What are teachers learning? *Education Policy Analysis Archives, 14*(5).

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education, 79*(4), 33-53.

Miller, J. W., McKenna, M. C., & McKenna, B. A. (1998). A comparison of alternatively and traditionally prepared teachers. Journal of Teacher Education, 49(3), 165-176.*

Monk, D. H., & King, J. A. (1994). Multilevel teacher resource effects in pupil performance in secondary mathematics and science: The case of teacher subject matter preparation. In R. G. Ehrenberg (Ed.), Choices and consequences: Contemporary policy issues in education (pp. 29-58). Ithaca, N.Y.: ILR Press.

Moss, P. A., Schultz, A. M., & Collins, K. M. (1998). An integrative approach to portfolio evaluation for teacher licensure. *Journal of Personnel Evaluation in Education, 12*(2), 139-161.

National Center for Education Statistics. (2004). *Digest of education statistics, 2004*. Retrieved January 2006, from http://nces.ed.gov/programs/digest/d04/lt1.asp#c1_1

National Research Council. (2001). *Testing teacher candidates: The role of licensure tests in improving teacher quality*. K. J. Mitchell, D. Z. Robinson, B. S. Plake, & K. T. Knowles, (Eds.). Washington, DC: National Academy Press.

National Research Council. (2008). *Assessing Accomplished Teaching: Advanced-Level Certification Programs*. M. D. Hakel, J. A. Koenig, & S. W. Elliott, (Eds.). Washington, DC: National Academy Press.

Pecheone, R. & Chung, R. (2007) *Technical report of the Performance Assessment for California Teachers (PACT): Summary of validity and reliability studies for the 2003-04 pilot year*. Stanford, CA: PACT Consortium.

Pecheone, R., & Stansbury, K. (1996). Connecting teacher assessment and school reform. *Elementary School Journal, 97*(2), 163-177.

Pecheone, R. L., Pigg, M. J., & Chung, R. R. (2005). Performance Assessment and Electronic Portfolios: Their Effect on Teacher Learning and Education. *The Clearing House, 78*(4), 164-176.

Podgursky, M. (2001). Defrocking the National Board: Will the imprimatur of "Board Certification" professionalize teaching? *Education Matters 1*(2).

Popham, W. J. (1990). Face validity: Siren song for teacher-testers. In *Assessment of Teaching: Purposes, Practices, and Implications for the Profession* (pp. 1-14). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Powers, D. E. (1992). *Assessing the Classroom Performance of Beginning Teachers: Educators' Appraisal of Proposed Evaluation Criteria* (No. RR-92-56). Princeton, NJ: Educational Testing Service.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 79*(2), 417-458.

Rosenfeld, M., & Kocher, G. K. (1998). *A job analysis of the content knowledge and skill areas important for newly licensed (certified) elementary school teachers (grades K-6)* (Research Report). Princeton, NJ: Educational Testing Service.

Selke, M., Mehigan, S., Fiene, J., & Victor, D. (2004). Validity of standardized teacher test scores for predicting beginning teacher performance. *Action in Teacher Education, 25*(4), 20-29.

STATA. (2005). *STATA for Windows*. College Station, TX: Stata Press.

Touchstone Applied Science Associates (TASA). (2005). *TASAtalk*. Retrieved January 1, 2006, from http://www.tasaliteracy.com/tasatalk/05-spring.pdf Touchstone Applied Science Associates (TASA). (2006). *Degrees of Reading Power*. Retrieved January 1, 2006, from http://www.tasaliteracy.com/drp/drp-main.html Vandevoort, L. G., Amrein-Beardsley, A., & Berliner, D. C. (2004). National board certified teachers and their students' achievement. *Education Policy Analysis Archives, 12*(46), 117.

Walsh, K. (2001). *Teacher certification reconsidered: Stumbling for quality*. Baltimore, MD: The Abell Foundation

Wayne, A., & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research*, *73*(1), 89-122.

Wenglinsky, H. (2002). How schools matter: The link between teacher classroom practices and student academic performance. *Education Policy Analysis Archives, 10*(12). Retrieved 6/20/09 from <http://epaa.asu.edu/epaa/v10n12/>http://epaa.asu.edu/epaa/v10n12/.

Wenglinsky, H. (2003). Using large-scale research to gauge the impact of instructional practices on student reading comprehension: An exploratory study. *Education Policy Analysis Archives, 11*(19), 1-19.

Wilson, S. M., Darling-Hammond, L., & Berry, B. (2001). *A case of Successful teaching policy: Connecticut's long-term efforts to improve teaching and learning* (No. R-01-2). Seattle, WA: Center for the Study of Teaching and Policy.

Youngs, P. (2002). *State and District Policy Related to Mentoring and New Teacher Induction in Connecticut* (No. ED472133). New York, NY: National Commission on Teaching & America's Future.