education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

epaa aape

Arizona State University

Volume 22 Number 6 February 10th , 2014 ISSN 1068-2341

Evaluating the Validity of Portfolio Assessments for Licensure Decisions

Mark Wilson
University of California, Berkeley

P.J. Hallam
California Department of Education

Raymond Pecheone
Stanford University
&

Pamela A. Moss
University of Michigan, Ann Arbor
USA

Citation: Wilson, M., Hallam, P J, Pecheone, R.L., Moss, P. A. (2014) Evaluating the Validity of Portfolio Assessments for Licensure Decisions. *Education Policy Analysis Archives, 22* (6) Retrieved [date], fromhttp://dx.doi.org/10.14507/epaa.v22n6.2014

Abstract: This study examines one part of a validity argument for portfolio assessments of teaching practice used as an indicator of teaching quality to inform a licensure decision. We investigate the relationship among portfolio assessment scores, a test of teacher knowledge (ETS's Praxis I and II), and changes in student achievement (on Touchstone's Degrees of Reading Power Test [DRP]). Key questions are the extent to which the assessment of teaching practice (a) predict gains in students' achievement and (b) contribute unique information to this prediction beyond what is contributed by the tests of teacher knowledge. The venue for our study is Connecticut State Department of

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 5/09/2011
Revisions received: 7/15/2013
Accepted: 8/28/2013

Education's (CSDE) support and licensure system for beginning teachers, the *Beginning Educator Support and Training* (BEST) program (as it was implemented at the time of our study). We investigated whether elementary teachers' mean effects on their students' reading achievement support the use of BEST elementary literacy portfolio scores as a measure of teaching quality for licensure, using a data set gathered from both State and two urban school district sources. The HLM findings indicate that BEST portfolio scores do indeed distinguish among teachers who were more and less successful in enhancing their students' achievement. An additional analysis indicated that the BEST portfolios add information that is not contained in the Praxis tests, and are more powerful predictors of teachers' contributions to student achievement gains.
Keywords: Teacher portfolio assessments, teacher standardized tests, correlational evidence of validity

La evaluación de la validez de las evaluaciones de *portfolios* para las acreditaciones
Resumen: Este estudio examina una parte de un argumento de validez para las evaluaciones de los cartera de la práctica docente utilizados como un indicador de la calidad docente para informar una decisión licencia. Investigamos la relación entre los resultados de la evaluación de cartera , una prueba de conocimientos del profesorado (Praxis I del ETS y II), y los cambios en los logros del estudiante (en grados de piedra de toque de la prueba eléctrica de la lectura [DRP]) . Las cuestiones clave son la medida en que la evaluación de la práctica docente (a) predecir las ganancias en el rendimiento de los estudiantes y (b) contribuir información única a esta predicción más allá de lo aportado por las pruebas de conocimiento de los maestros . El lugar elegido para nuestro estudio es el Departamento de apoyo y otorgamiento de licencias (CSDE) Sistema de Educación para los maestros principiantes del Estado de Connecticut , el Principio Educador , y la programa (BEST) (como se implementó en el momento de nuestro estudio) Adiestramiento. Hemos investigado si los "efectos medias en sus alumnos los profesores elementales logros en lectura apoyan el uso de las mejores puntuaciones de la cartera de alfabetización primaria como una medida de calidad de la enseñanza para obtener la licencia , el uso de un conjunto de datos recogidos de Estado y de dos fuentes distrito escolar urbano . Los hallazgos indican que las puntuaciones HLM mejor cartera de hecho distinguen entre los profesores que se encontraban cada vez menos éxito en la mejora de los logros de sus alumnos. Un análisis adicional indicó que los mejores carteras añadir información que no está contenida en las pruebas Praxis , y son más potentes predictores de las contribuciones de los profesores a las ganancias de rendimiento de los estudiantes .
Palabras clave: evaluación de la cartera de los maestros , las pruebas estandarizadas de maestros, pruebas de correlación de validez

Avaliar a validade da avaliação das carteiras de acreditação
Resumo: O presente estudo analisa uma parte de um argumento de validade para avaliação da carteira de prática de ensino utilizadas como um indicador da qualidade do ensino para informar a decisão de licença. Nós investigamos a relação entre os resultados do portfólio de avaliação, um teste de conhecimentos e competências (ETS Praxis I e II) , e as mudanças no desempenho do aluno (em graus pedra de toque de teste elétrico leitura [DRP]) . As questões-chave são a medida em que a avaliação da prática docente (a) prever ganhos no desempenho dos alunos e (b) contribuir com informações exclusivas para a predição além da evidência fornecida pelo conhecimento dos professores. O local escolhido para o nosso estudo é o suporte Departamento e licenciamento (CSDE) Sistema de Educação para professores iniciantes do Estado de Connecticut, o Educador Começando , eo programa (BEST) (como implementado no momento do nosso estudo) Treinamento. Nós investigamos se os "efeitos médios em seus professores elementares de leitura os alunos realização apoiar o uso das melhores pontuações da carteira de alfabetização primária como

uma medida da qualidade da educação para o licenciamento , o uso de um conjunto de dados coletados fontes do Estado e dois distrito escolar urbano . os resultados indicam que as pontuações melhor portfolio HLM realmente distinguir entre os professores que estavam menos bem sucedido de sempre na melhoria da realização dos seus alunos. Uma análise indicou ainda que os melhores portfolios adicionar informações não contidas nos testes Praxis , e são preditores mais poderosos de contribuições dos professores para os ganhos de desempenho dos alunos .
Palavras-chave: avaliação de portfólio para os professores , testes padronizados, os professores testes de correlação de validade

Evaluating the Validity of Teaching Portfolios for Licensure Decisions[1]

The authority to grant a license to teach in this country resides with individual states. Each state has a licensure system in place intended to ensure that children are taught by competent teachers. The decision by a state to grant a license to teach is typically based upon multiple pieces of evidence from different stages of a prospective teacher's preparation and induction into the profession. While the pattern of evidence varies considerably from state to state, the following sources of evidence are among those most conventionally used: graduation from an accredited teacher preparation institution; successful completion of practice teaching; and passing one or more tests of basic skills, content knowledge, and/or pedagogical knowledge, and (in a very few cases) assessment of actual teaching practice.

There is a growing call for evidence of teaching practice in licensure decisions and in teacher evaluation more generally. The National Research Council (NRC) (2001), in its review of teacher licensure tests, called for indicators that go beyond testing to include "assessments of teaching performance in the classroom, of candidates' ability to work effectively with students with diverse learning needs and cultural backgrounds and in a variety of settings, and of competencies that more directly relate to student learning" (NRC, 2001, p 172). A growing number of reports from RAND (Ball and RAND Mathematics Study Panel, 2003) and from the National Academy of Education (Darling Hammond and Baratz-Snowdon, 2005; Darling-Hammond and Bransford, 2005) echo this call for assessment of knowledge in use in teaching practice. Federal legislation in the past decade, especially the 2001 "No Child Left Behind" (NCLB) Act and the "Race to the Top" (RTT) Program funded by the 2009 American Recovery and Reinvestment Act have spurred attention by tying funding to teacher evaluation—NCLB by calling on states to identify highly qualified teachers and RTT by calling more specifically for evaluations that link teachers to their students' achievement. Major projects by NRC (2008) and the Gates-sponsored Measures of Effective Teaching (MET 2013, 2013) address questions about the role of portfolios and observation systems,

[1] 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. 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, U.S. Department of Education, through Grant #R305T010511 to the University of Michigan. The opinions, findings, and recommendations expressed are those of the authors and do not represent views the Institute of Education Science or the U.S. Department of Education.

respectively, in evaluating teaching quality. These studies focus on the practice of experienced teachers.

Educators and policy makers at the state level are facing decisions about whether to include direct measures of teacher practice, and, if so, what measures of practice to include in the design of their licensure systems. This article contributes to informing state policy decisions around assessing and supporting effective teaching. Prominent measures of teaching practice used at scale include observations systems, both live and video-based, and teacher prepared portfolios or exhibits. Such assessments are resource intensive, entailing the development of technology-based systems for documenting practice, and also involving extended teacher time for preparing the assessment, as well as time to train qualified scorers. Key questions are the extent to which the assessments of teaching practice (a) predict gains in students' achievement and (b) contribute unique information to this prediction beyond what is contributed by other less resource intensive measures of teaching quality, especially the on-demand tests of teacher knowledge that have been widely used to date. That is the issue to which this study contributes in the context of a state-sponsored portfolio assessment system.

The venue for our study was Connecticut State Department of Education's (CSDE) induction and licensure system for beginning teachers, *Beginning Educator Support and Training* (BEST). The BEST portfolio assessments were used to support second stage licensure decisions for beginning teachers in their second or third year of teaching**.** The BEST portfolio system was shaped by the assessment design of the National Board for Professional Teacher Standards (NBPTS) certification system, including the content specific focus used by the NBPTS. The BEST portfolio assessments required early career teachers to prepare a portfolio of their teaching practice focused on a unit of instruction that included: goals and lesson plans for the unit, instructional artifacts, video tapes of teaching, samples of evaluated student work, and commentary and reflection on their practice. Trained raters evaluated these in key competency areas against specific state teaching standards. These data were the basis for a decision about whether the teacher had met the performance standards to be granted a renewable professional license.

We focus, in particular, on the BEST elementary literacy portfolio. Specifically, we investigate the relationship among scores on the literacy portion of the BEST portfolio assessment, ETS's PRAXIS I and II Tests of teachers' basic skills and pedagogical knowledge, and students' gains on Touchstone's Degrees of Reading Power (DRP) Tests from two large urban districts. We consider first the extent to which the pattern of relationships among these measures supports the validity of the portfolio assessment as an indicator of teaching quality. Then we use a hierarchical linear modeling (HLM) analysis (Bryk & Raudenbush, 1988; Raudenbush & Bryk, 2002) to consider the extent to which the scores from the portfolio assessment contribute unique information to the prediction of student achievement gains. As such, this study provides criterion related validity evidence for the portfolio assessment and the other sources of evidence contributing to the licensure decision, treating student achievement gains as the criterion measure.

We note, as of this writing, the BEST portfolio assessment system policy has been changed based on recent legislation. The assessment structure and tasks have beed modified with a greater emphasis on mentoring. However, the newly developed BEST assessment continues to be a measure of teaching that leads to a professional licensing decision in Connecticut. Moreover, the information from our study is particularly pertinent in light of (a) the development of the Performance Assessment for California Teachers (PACT; Author, 2005), and now a national pilot of performance assessments (edTPA, formerly Teacher Performance Assessment) currently underway in 26 states for both of which the Connecticut BEST Program was a progenitor. The national pilot uses portfolio-like performance assessments developed by the Stanford Center for

Assessment, Learning, and Equity (SCALE),[2] in partnership with the American Association of Teacher Educators (AACTE), and faculty design teams from participating states. To help manage the edTPA at scale the Pearson Corporation is the designated operations partner.

In the sections that follow, we first provide a brief literature review of those studies that have examined the relationship between assessments of teaching practice and gains in student achievement with attention to the methodological hurdles these studies addressed, so readers can compare our findings to those reported there. Then we describe in detail the context of the research, the sources of evidence on which the licensure decision is based, the validity evidence available for our key predictor and criterion variables, and the rationale for and limitations of the study design in light of our research question. Our results section includes descriptive information for each variable, preliminary examination of what the relationship among the variables contributes to our understanding of the validity of the portfolio assessment as a measure of teaching quality, and finally the HLM analysis which addresses the key question of what the portfolio assessment contributes to the prediction of student achievement gains beyond what the other measures contribute. As readers will see, the HLM findings indicate that BEST portfolio scores do indeed distinguish among teachers who were more and less successful in enhancing their students' achievement. The analysis indicated that the BEST portfolios add information that is not contained in the Praxis tests, and are more powerful predictors of teachers' contributions to student achievement gains. In the concluding sections, we situate our findings amongst those emerging from similar studies with experienced teachers, using observations (MET, 2011, 2012) and portfolios (NRC, 2008). We make recommendations for next steps in this research agenda relevant to licensure systems. We also suggest questions that educators and policy makers responsible for designing licensure systems might want to consider in choosing among different approaches to the assessment of teaching practice and the research agenda this implies.

Relationships Among Assessments of Teaching Practice and Students' Achievement as Evidence of Validity

Our study focuses on one particularly crucial aspect of validity evidence that until recently (NRC, 2001, 2008) was not routinely available for assessments of teaching quality, whether practice based, paper and pencil, or an administrative proxy (i.e., credentials of various sorts): the relationships between measures of teaching quality and student achievement gains. As the authors of the 2012 MET report, "*Ensuring Fair and Reliable Measures of Effective Teaching*" note, "Teachers shouldn't be asked to expend effort to improve something that doesn't help them achieve better outcomes for their students. If a measure is to be included in formal evaluation, then it should be shown that teachers who perform better on that measure are generally more effective in improving student outcomes" (p. 15). This study focuses on what the MET authors describe as this "central" test of validity. In this section we review other studies that have explored this relationship, first with what might be described as proxies of teaching practice (e.g., credentials, paper and pencil tests), second with assessments involving portfolios or exhibits prepared by teachers (which is the focus of our study), and finally with observation instruments (which was the focus on the MET study). For policy makers who wish to include a measure of teaching practice in their licensure policy, or in their teaching evaluation policy more generally, choices among such measures will need to be made. This review will also show the relatively unique contribution our study makes to the literature on direct assessment of teaching practice which has tended to emphasize assessments of experienced teachers rather than beginning teachers.

[2] Co-author is the leader of this effort for SCALE and was the leader of the effort to develop the BEST system.

Studies Involving "Proxies" of Teaching Practice

In general, statistically significant and important findings are often difficult to achieve in research on the relationship between teacher characteristics and student achievement. Milanowski (2004), in his study of the relationship between teacher evaluation scores and student achievement, 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). Wenglinsky (2002) studied relationships among teacher characteristics and student academic performance by applying multilevel modeling to the 1996 National Assessment of Educational Progress in mathematics and concluded, "Like most of the prior research, this model finds no significant relationship to test scores for most of the characteristics, with the exception of the teacher's college-level coursework as measured by major or minor in the relevant field." Similarly, Glass (2002) concluded that traditional psychometric techniques such as using scores from ability, achievement, other paper-and-pencil tests, and GPAs to predict teaching effectiveness in terms of student achievement have failed.

Studies that do report relationships between student achievement and teacher characteristics are often hotly debated. For example, several studies on the impact of certification reported evidence that found higher achievement for students of teachers from traditional routes than those from alternative routes, and for fully certified teachers (as opposed to partially-certified teachers) (Darling-Hammond, 2000; Darling-Hammond, 2001b; Goldhaber & Brewer, 2000; Hawk, Coble & Swanson,1985; Laczko-Kerr, Berliner, 2002; Miller, McKenna & McKenna, 1998; Monk & King, 1994). Conversely, a 2001 review by Walsh of approximately 150 studies on teacher licensure asserted that many studies did not provide evidence that students taught by uncertified teachers performed any better or worse than those of certified teachers. Walsh's publication touched off a heated debate about the quality and interpretation of research on teacher effectiveness (Darling-Hammond, 2001a) and increased attention to concerns about educational study methodologies (Ballou & Podgursky 1999). Despite subsequent studies (Clotfelter, Ladd, & Vigdor, 2007; Rivkin, Hanushek & Kain, 2005; Wayne & Youngs, 2003), these controversies have left the field still searching for clear conclusions.

These studies suggest that these administrative proxies of teaching practice are not of sufficient validity for documenting the quality of teaching. Studies involving the relationship of paper and pencil tests to gains in student achievement measures, while rare, show mixed results

A study of Praxis I and Praxis II by Goldhaber (2007) did find a weak positive relationship between some Praxis tests and student achievement. Teachers who met North Carolina's Praxis II requirements were somewhat more effective in math and reading[3]. Further, the higher teachers scored on the Praxis *Curriculum, Instruction & Assessment* (CIA*)* test, 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 (Clotfelter, Ladd, & Vigdor 2006), 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 students over time. In interpreting the results, the authors raised the concern that the nonrandom sorting of teachers did have an impact on the estimated relationship between teacher test performance and student achievement. A more

[3] The differences are quite small, however: In comparing the top quintile to the bottom quintile, the difference for reading was about 2.4 percent of a standard deviation, and 3 percent for math.

recent study undertaken as part of the MET (2013) project showed no significant relationships between student achievements gains and content knowledge for teaching tests in English Language Arts and Mathematics. They suggested that the measures were still early in development and that the lack of significant relationship may be due to technical issues that will be resolved as the assessment is further developed.

Studies Involving Portfolios or Exhibits of Practice Prepared by Teachers

These assessments provide more direct measures of teaching practice than the proxies describe above. Portfolios and exhibits usually involve multi-media records of practice selected by teachers to represent their practice, along with extended commentary that situates, provides a rationale for, and reflects on that practice in response to standardized guidelines. These sorts of assessments differ from the observation systems described below by giving the teacher considerable control over the timing and focus of the recordings of their practice and opportunity for extended commentary.

Much of the relevant work here has focused on the assessments of the National Board for Professional Teaching Standards, a certification process designed to identify accomplished teaching (Bond, Smith, Baker, & Hattie, 2000; Cavaluzzo, 2004; Goldhaber & Anthony, 2004; Ladson-Billings & Darling-Hammond, 2000; Lustick & Sykes, 2006; Vandevoort, Amrein-Beardsley, & Berliner, 2004). National Board Assessments involve two major components—a teaching portfolio and an on-demand timed assessment. The results reported here treat the assessment as a whole without distinguishing among the components. The extent to which completed NBPTS studies definitively indicate that the students of National Board certified teachers achieve significantly higher academic gains (compared to the students of other teachers) has been debated, and the results have been mixed (Bond, 2001; Cunningham & Stone, 2005; Podgursky, 2001), highlighting a need for further exploration of the relationship between student achievement and teacher portfolio assessment. A recently completed study by the National Research Council (2008) reviewed 10 such studies.[4] Of these, they found seven studies with sufficient sample size and methodological sophistication to allow sound conclusions about the observed relationships. The NRC report highlighted the sorts of methodological problems the authors of these studies faced, including non-random assignment of students to teachers and teachers to schools (which made it harder to distinguish teaching quality from other factors that might impact the relationship) and the nesting of students within classrooms (which lead to effect estimates biased in favor of statistical significance). In the studies they considered methodologically sound, these problems were addressed through statistical controls at the individual, classroom, and/or school level[5] and through multi-level models or other statistical correction procedures[6] that took nesting into account. Only one study the panel reviewed actually involved within-school random assignment for teachers. While some studies compared Board Certified Teachers to non-Board Certified Teachers, the report's authors noted that the stronger studies distinguished comparison groups between those who had applied for board certification but not attained it and those who had never applied. They concluded:

> "Studies that compared test score gains for students of teachers who were and were not successful in earning board certification consistently found statistically significant differences between the two groups. Results from comparisons of test score gains for students of board-certified teachers and nonapplicants were less consistent" (p. 171).

[4] An eleventh study they located focused on an alternative student outcome.

[5] As detailed more fully in the report, these included covariates as well as fixed effects models at the student, teacher, and school levels.

[6] These statistical corrections procedures estimate "robust standard errors" (NRC, p. 161).

The NRC panel then commissioned two teams of researchers to re-analyze the data sets from two states (North Carolina and Florida) they considered most robust comparing alternative models for estimating the relationship. The comparisons showed the findings were more sensitive to the state context than to model specification (p. 172). The results for the model they considered to be the strongest[7] were described as follows:

> "Compared with other teachers, board certified teachers in North Carolina raise test scores about 7 percent of a standard deviation more in math and 4 percent of a standard deviation more in reading. In Florida, board certification is associated with a smaller increase of about 1 percent of a standard deviation in mathematics and about 2 percent of a stand deviation in reading. The coefficients for Florida were not statistically significant." (p. 173).

Their findings led them to conclude that while the differences are small (and not entirely consistent), national board certification distinguishes more effective teachers from less effective teachers with respect to student achievement in substantively meaningful ways. (p. 179).

While the studies involving National Board Certification focused on experienced teachers, one small study explored the relationship between practice assessments from preservice teachers and subsequent student achievement during their first years of teaching (Newton, 2010). [We note that this study was exploratory in nature and would not likely have passed the criterion the NRC panel used to distinguish the seven studies that warranted conclusions about relationships.] The practice assessments, prepared by pre-service teachers, were part of the Performance Assessment for California Teachers (PACT) and consisted of portfolio based assessment tasks patterned after the BEST (and NBPTS) assessment tasks. The study examined the relationship between PACT scores for 14 teachers in grades 3-6 in one district and the teachers' subsequent teaching effectiveness estimated by their students' gain scores (n=259) on a standardized ELA achievement test. Newton reported "total PACT score correlated approximately .50 with teacher value-added….For each additional point a teacher scored on PACT (evaluated on a 4 point scale), his/her students averaged a gain of one percentile point per year on the California Standards Tests as compared with similar students." While the focus of this study most closely resembles our own, our sample size is considerably larger and our methodology thus able to address the concerns with nesting and non-random assignment in ways consistent with studies relied on by the NRC panel.

Studies Involving Observation Systems

As we noted above, observations systems typically differ from the sorts of portfolio assessment that is the focus of our study by allowing teachers less control over when and what is observed, little or no opportunity to examine teachers' responses to students' written work, and less opportunity for commentary.. However, the observation systems are typically far less time consuming for teachers (a tradeoff to which we'll return in our conclusion). As we write, the Measures of Effective Teaching Project (metproject.org) has completed a multi-year study examining the relationship between various measures of teaching quality and student achievement gains with nearly 3000 volunteer teachers. Of the reported findings of the MET study, our focus is on the 2012 and 2013 reports. The study reports focused on a sample of 1,333 teachers who taught ELA or Math in Grades 4-8 and agreed to be randomly assigned to classes within schools for the final year of the study. They considered five observation systems that focused on instructively different aspects of teaching, including those that were more generic and more subject-specific; those that focused on :

- Framework for Teaching (or FFT, developed by Charlotte Danielson of the Danielson

[7] This model "used the gain score as the outcome measure and estimated both student and school fixed effects" (p. 172).

Group),

- Classroom Assessment Scoring System (or CLASS, developed by Robert Pianta, Karen La Paro, and Bridget Hamre at the University of Virginia),
- Protocol for Language Arts Teaching Observations (or PLATO, developed by Pam Grossman at Stanford University),
- Mathematical Quality of Instruction (or MQI, developed by Heather Hill of Harvard University), and
- UTeach Teacher Observation Protocol (or UTOP, developed by Michael Marder and Candace Walkington at the University of Texas-Austin**).** (MET, 2012, p. 2)

Each participating teacher provided multiple videos, all of which were scored by trained raters in at least three of the observation systems named above. The authors concluded that "all five of the observations were positively associated with student achievement gains." (MET, 2013, p. 6), including both gains on state administered standardized achievement tests and specially administered tests that addressed more conceptual understanding in math and short essay responses in writing. In addition, the authors considered the predictive power of previous years value added scores (with a different class of students) and students' ratings of teachers' classroom practice. They noted that "combining observation scores with evidence of student achievement gains and student feedback improve predicative power" (MET 2012, p. 9). Consistent with the findings reported above on proxies, the authors noted that "in contrast to teaching experience and graduate degrees, the combined measure identifies teachers with larger gains on state tests" (p. 12). The analyses released in 2012 addressed the problems of non-random assignment and nesting with multi-level modeling and statistical controls as had the authors of the studies reviewed by the NRC panel. In 2013, they released results of additional analyses of these teachers who had been randomly assigned within schools for the last year of the study. Their findings were similar to the previous year's findings and allowed them to conclude that "the adjusted measures [from the previous year with non-random assignment] did identify teachers who produced higher and lower achievement gains following random assignment (MET, 2013, p. 5) suggesting that the statistical controls had been effective and could be used when random assignment was not feasible (as is routinely the case). We'll draw on these findings in our conclusion, as they bear on the sorts of choices policy makers face in designing licensure systems or systems of teacher evaluation more generally.

Methods: Instruments, Data Sources and Analyses

The CSDE'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 be eligible for an initial license, prospective teachers had to pass appropriate Praxis tests (i.e., PRAXIS I and PRAXIS II as well as fulfill other program requirements); to be eligible for a provisional license teachers were required to successfully complete the BEST program, including passing the BEST portfolio assessment; and, finally, to be eligible for a professional license teachers had to meet state level requirements for Continuing Education Units (CEUs) as well as fulfill additional professional requirements (e.g., Masters degree). The BEST program was a two to three year comprehensive program of support and assessment. The support component consisted of individual mentors or support teams from the teachers' own school or district, who successfully participated in state sponsored support training.

The portfolio assessment component of the BEST program 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), since the participants were 3rd through 6th grade multiple-subject teachers (CSDE, 2006). EE portfolios required teachers to document five to eight hours of instruction on one literacy unit and one mathematics unit for one class of students. Documentation included 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[8].

In the BEST program, beginning teachers were required to demonstrate, through the portfolio assessment, acceptable levels of essential teaching competencies related to four domains of teaching: (a) instructional design, (b) instructional implementation, (c) assessment of learning, and (d) analyzing teaching and 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 teaching. For the purposes of this study, each teacher's first official submission and the associated BEST score was used in data analyses.

As implemented at the time of our study[9], the portfolios were evaluated by experienced teachers who have received at least five days of training at a regional training center 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 assessor was called in to reconcile the scores. Assessors first took notes on the portfolio based upon a series of guiding questions[10] (GQ's) and associated rubrics also provided to the beginning teacher. The questions were organized into four categories: instructional design (3 GQ's), instructional implementation (planning) (7 GQ's), assessment of learning (5 GQ's) , and analyzing teaching and learning (2 GQ's). Then assessors decided on one of four performance levels based upon an integrative holistic scoring rubric that described the performance levels[11]. Assessors reviewed their notes and cited evidence for each guiding question to arrive at a score. They also completed a "feedback rubric" which contained performance level descriptions on a four point scale for each guiding question and this was used to give more specific feedback for the beginning teacher. A sample of portfolio notes that provided evidence for each GQ rubric score were audited by an assessor trainer who provided additional training if assessors seemed to be drifting off calibration. Independent re-evaluations were conducted for all failing portfolios, as well as for a sample of just-passing portfolios (i.e., 2 on a 4 point scale), and for any portfolios where the trainer did not feel the documented evidences justified the score given. The level of inter-rater agreement for each guiding question was evaluated on the basis of the percent of exact and adjacent scores. Rubric scores that differed by plus or minus 2 points were judged to be unreliable and triggered a third independent evaluation of the portfolio. Assessors were expected to score approximately 2 to 3 portfolios per day. All beginning teachers received a feedback report that highlighted their performance on each of the 17 guiding questions in order to provide teachers with an analytic profile of their strengths and weaknesses. Mentors received specialized training on interpreting the feedback report, which included strategies to both build on teacher strengths and address areas of weakness. Reliability information was routinely maintained based upon the initial scoring by two assessors and the independent audited rescores. Pecheone & Stansbury (1996) and Youngs (2002) indicated that the inter-rater reliability coefficients for the portfolios were at acceptable levels (r = .72 to .76).

---

[8] 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.

[9] See Appendix A for an overview of the BEST Portfolio scoring process.

[10] See Appendix B for a list of the guiding questions.

[11] See Appendix C for the Decision Guide for the Holistic Evaluation.

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 became official, the state conducted a special reliability study where a sample of portfolios was scored by multiple pairs of readers including an independent audit of portfolios around the cut-score. Alignment among standards, portfolio handbooks, scoring materials, and training procedures was also investigated. Validity studies of external relationships involving BEST portfolio scores had not been conducted at the time of this writing.

The CSDE's Use of *Praxis* Tests

CSDE provided data on both Praxis I and Praxis II tests for use in this study. The Praxis Tests were developed and scored by the Educational Testing Service (ETS). 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 either 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, and 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 2008). 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.

Regarding technical quality of the Praxis Series, a wide-ranging review conducted under the auspices of the National Research Council concluded that the evidence collected on the use of 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" (Mitchell, Robinson, Plake, & Knowles, 2001). However, the NRC review did not find any evidence at the time of the relationship between student achievement and teacher performance on either the Praxis I or Praxis II tests. The one study by Goldhaber (2007) described above provides some criterion related evidence of the relationships between Praxis tests and gains in student achievement.

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* (DRP; 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 (Connecticut State Department of Education, 2006; Touchstone Applied Science Associates, 2006). Findings from study of the psychometric properties of the DRP indicated that it has high level of reliability (test-retest = .95) and other aspects of technical quality were deemed adequate for the recommended uses of the test (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. Of course, like all standardized tests, DRP has limitations in terms of its validity as a measure of true student ability—however, as such tests are the only available comparable measures, we use this one in our study.

The Data Set

The data set constructed for this study was originally collected by State and district agencies: The data are not self-report, but are "official data" gathered from government records. The design is somewhat complicated in that it is a combination of a state-level data set (i.e., the data about the teachers), and two school district-level data sets (i.e., the data about the students). Thus, while at the first level it is a sample of school districts, at the second level, it is a census of all the teachers (and their associated data) within those districts that fit our profile and whose data were available.

Two urban Connecticut districts were selected on the basis of (a) 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, and (b) their willingness to allow data to be used for this project. The availability of such data was a crucial aspect governing the potential success of this project. A superior design would involve randomization among teachers and students, but this was simply not feasible. Information about teachers and their students was collected under approved guidelines of the Institutional Review Boards at the home universities of the principal investigators of the project, and following the guidelines for the CSDE as well as the two school districts.

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 (gender, ethnicity, district and grade level). 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.

Tables 1 and 2 provide 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. As would be expected for teachers in urban districts, they have, as a group, higher percentages of Hispanics and African Americans than the state as a whole. The sample included 61 teachers in District 1 and 43 in District 2. The teachers were fairly evenly spread across Grades 3 through 6.

Table 1

2002 to 2005 Teachers' Gender and Ethnicity[a]

Gender				Ethnicity			
	n	%	% in CT		n	%	% in CT
male	16	15	35	African Am.	12	12	3
female	87	84	74	Euro. Am.	75	72	93
NA[b]	1	1		Hispanic	14	13	3
total	104			Other	2	2	
				total	104		

[a] Note: percentages may not always add to 100 because of rounding.
[b] Note: NA indicates "not available."

Table 2
2002 to 2005 Teachers' District and Grade Levels[a]

District			Grade		
	n	%		n	%
1	61	59	3	20	19
2	43	41	4	37	36
total	104		5	21	20
			6	23	22
			NA	3	3
			total	104	

[a] Note: percentages may not always add to 100 because of rounding.
[b] Note: NA indicates "not available."

The student data were provided by the two school districts. There were 1041 male students (51%) and 961 female students (49%). The results in Table 3 indicate that almost half of the students were African American, 49%, with Hispanic being the next largest percentage, 37%. Almost all of the students qualified for free or reduced lunch, 94%, indicating that the students in this study are from high poverty families. The percentages of African American and Hispanic students, and students taking free or reduced lunch, is higher than for the rest of the state. Table 4 shows information regarding students' special education and English Language Learner (ELL) status. The percentage of students that qualified for special education is 11% (approximately the same as for the state as a whole), and 13% qualified for ELL services (a bit more than twice the percentage for the state as a whole). 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 3
2002 to 2005 Students' Ethnicity and Lunch Status[a]

Ethnicity				Lunch			
	n	%	% in CT		n	%	% in CT
Native Am.	77	4	0	Free/			
Asian Am.	15	1	3	reduced	1622	94	26
African Am.	974	49	14	Full	86	4	4
European Am.	132	6	68	NA	20	1	
Hispanic	736	37	14	Total	1977		
Other	41	2					
NA[b]	2	.1					
total	1977						

[a] Note: percentages may not always add to 100 because of rounding.
[b] Note: NA indicates "not available."

Table 4
2002 to 2005 Students' Special Education Status, and English Language Learners[a]

Special Ed				ELL			
	n	%	% in CT		n	%	% in CT
yes	221	11	11	yes	255	13	5
no	1386	70	89	no	1358	69	95
NA	370	19		NA	364	18	
total	1977			total	1977		

[a] Note: percentages may not always add to 100 because of rounding.
[b] Note: NA indicates "not available."

Covariates

Absent a randomized design for data collection, one needs to control for as many potentially confounding variables as possible and typically the way to do this is to include these variables in the analysis as covariates, at either the student level or the teacher level. Given that the purpose of the study is to seek evidence testing the sensitivity of an instrument (the BEST portfolio scores) to aspects of teacher quality, it seems inappropriate to control for teacher variables as covariates. Hence, we concentrate on student-level variables in these analyses (but we did carry out some exploratory investigations regarding teacher covariates).

Regarding student covariates, an initial list of covariates was generated from our search of the literature and that helped us identify the most likely candidates from the set of covariates available to us. At the student level, students' socio-economic status is consistently found to be a factor in student achievement. In this data 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 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. We decided that where there was very little missing data (1 or 2 cases), that we would code those entries as "missing." But, for variables with greater amounts

of missing data, in order to check whether the missing data was possibly influential, we included a separate "missing data" variable for each such covariate (i.e., 1 for "missing," 0 for "not missing").

Correlational Analyses

Three correlational analyses were completed using traditional Pearson correlation coefficients (with statistical significance evaluated using a two-tailed alternative). The first analysis correlated BEST portfolio scores, Praxis I scores, and Praxis II scores with student gain scores. The second correlated BEST portfolio scores with Praxis I and II scores. The third analysis used partial correlations, holding the pre-test scores constant to correlate student post-test scores with (a) portfolio scores and (b) Praxis II scores.

In interpreting these findings one must keep in mind the original purpose of the Praxis Tests: They are focused on identifying those teacher candidates who possess the minimum knowledge, skills, and abilities necessary to work as entry-level teachers. In addition, they were not designed to identify outstanding teachers with higher performance scores or to be used to rank order teachers based on their performance. Thus, the patterns in the findings observed is not surprising. In our reading of the literature, we generally agree with the NRC report cited earlier (NRC, 2001), in that the literature does support the psychometric soundness of the series, but we do note that there has been more recent work that does indicated some support for the link between them and student performance.

Hierarchical Linear Modeling (HLM)

Findings on the relationships between teacher characteristics and student achievement have been influenced greatly by advancements in methodologies for analyzing teacher characteristics. As well as examining the correlation coefficients, this study utilizes hierarchical linear modeling (HLM; (Raudenbush & Bryk, 2002) because it can help sort out the magnitude of impacts at different levels of the education system from which improvements in student learning emerge – in this case, the student and the teacher..

Although the idea of gain scores is intuitively appealing and a more straightforward method to explain to many audiences, it is often preferred to use the post-test scores as the outcome, with the pre-test scores as a covariate.

A 2-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 (i.e., pre-test scores on the DRP), 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 in the data set include teacher demographic data (such as gender), type of mentoring program, and prestige of undergraduate institution. Additional analyses indicated that none of the teacher level covariates (including Praxis scores) had statistically significant effects (at the standard $a=0.05$ level), which is consistent with the correlational results, so they are not discussed in the "Results" section below.

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 overall by weighting the more reliable data more heavily—effectively, this means that, for instance, the data for teachers with more students in their class will be weighted more heavily. This estimation technique is preferable to ordinary least squares estimates of residuals especially for this study because, indeed, 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 was chosen so that the results can be presented in terms of DRP Units, which have useful interpretability.

As there was missing data shown in Tables 3 and 4 at the student level, we included a missing data category as well for each variable with missing data. The reasons for these missing data are not known to us, as we can only report the governmental data that was available to us. However, including them as a separate code allows us to gauge whether their presence affects the basic findings. Note that we did not attempt this for variables that had very small amounts of missing data (i.e., 1 or 2 cases).

The interesting alternate approach described by Goldhaber and Anthony (2004) and Clotfelter et al (2006) uses fixed effects to try and control for the effect of teacher sorting evidenced by a positive correlation between initial student achievement and teacher scores. In this data set, the correlation between these variables is negative, -0.102 ($p= 0.3008$), revealing that the phenomenon observed by these researchers is not indicated for this data set—hence, we will use the more straightforward HLM approach.

Limitations of the Study

There are several limitations to this study that need to be borne in mind when interpreting the results. For one, this study is based on a secondary data analysis. The data were originally collected for other purposes, and then linked for the purposes of this study. Hence, there was no opportunity to apply randomization of any kind to strengthen the design. This also means that the staff of the study could not supervise the original data collection. Given the high stakes associated with many of these assessments, we believe that we can trust in the state's strategies for consistent data collection. Nevertheless, given the strictures of using data from a state-run licensure program, the project did undertake stringent and exhaustive 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 (see Table 6), and see this as helping sensitize the results to this issue. Third, the logistical difficulties of documenting performance indicators may be contributing factors. For example, Pecheone et al. (2005) noted that the potential for bias in the selection of artifacts from the portfolio assessment as evidence of teacher abilities is questionable because teachers know they will be evaluated on the basis of these artifacts. Portfolios cannot be taken as evidence of typical practice, but rather are more likely evidence of what teachers' consider to be their best practice. Other means of collecting data that allow us to document teacher knowledge and skills would strengthen the evidence on teacher learning. Finally, the representativeness of the student sample needs to be considered—the two school districts that were selected serve low SES areas, so the results should be seen in that light.

Results

Student Achievement

Overall, the data indicate that, for the students in this sample, achievement in reading comprehension is in general somewhat low but 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 covered a wide range, with 27 students at the lowest possible score of 15, to a high of 95. The students' mean posttest score was 44, which is in the expected range of 3rd

grade scores (recall that the sample includes students from grade 3 to 6). The 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 "Elementary School Textbooks" with books such as *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. 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).

Correlation Results

The correlations among student mean gain scores (averaged for each teacher), the overall scores on the BEST portfolios, and the Praxis I and II test scores are displayed in Table 5. Overall, the results are similar to the findings reported in the literature—without any controls for potential sources of bias, the correlation coefficients are low and not statistically significant. Results for partial correlations, controlling pretest scores were also calculated (but are not shown)—the general finding for these is the same as that for the simple correlations. Findings from the correlation analysis of BEST portfolio scores and Praxis scores are presented in Table 6. Again, these correlations are small and not statistically significant. Results for partial correlations, controlling for fall DRP scores, were also calculated—the general findings for these are the same as that for the gain scores. Specifically, the small and statistically non-significant correlations indicate that the portfolio scores are not related to the three standardized tests of teacher knowledge. This is not unexpected as the former is aimed at in-service accomplishment, whereas the latter are aimed at (various levels of) entry-level qualification.

Table 5
Correlations of Teacher Assessments and Mean Student Gain Scores

Assessment	Correlation with Mean Student Gain Scores	p-value	N
Portfolio Literacy	.16	.11	104
Praxis I Mean	-.03	.8	69
Praxis II (CIA)	-.02	.84	95
Praxis II (CAE)	.08	.44	92

Table 6
Correlations of Teachers' Portfolio Scores and Praxis Scores

Assessment	Correlation with Praxis Scores	P-value	N
Praxis I Mean	-.15	.22	69
Praxis II (CIA)	-.11	.29	95
Praxis II (CAE)	.08	.46	92

The outcome variable in our HLM analyses is DRP post score, with DRP pretest score always included as a student covariate. Table 7 indicates that seven of the student covariates were statistically significant. The most highly significant covariate was DRP pretest scores ($z = 26.56$; $p < 0.001$), which would be expected. The next six covariates 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—one could speculate as to why they are statistically significant, but the reasons for the missing status are not available to us. Nevertheless, it is important to note that, by including them in the analysis, we have supported the interpretation of the other coefficients as being robust to the missing data.

Table 7.
The Portfolio Model for the Urban School Districts

	Covariates	Coef.	SE	z	
Student Level					
	Pre DRP	0.47	0.02	26.56	***
	Grade	1.30	0.39	3.31	**
	ELL	-2.12	0.70	-3.02	**
	ELL miss.	-2.97	1.05	-2.82	**
	Female	0.22	0.37	0.59	
	African Am.	-1.03	0.92	-1.12	
	European Am.	1.98	1.15	1.71	+
	Hispanic	-0.57	0.98	-0.59	
	Lunch Status	-1.77	0.51	-3.44	**
	Lunch Status miss.	-2.10	2.00	-1.05	
	Special Ed.	-3.65	0.69	-5.30	***
	Special Ed. miss.	-4.47	1.02	-4.40	***
Teacher Level					
	Portfolio score	2.20	0.66	3.33	**
			ICC: 0.18		
			R^2_B: 0.80		
			R^2_W: 0.32		

[1]Statistical significance codes: + $p < 0.1$, * $p < 0.5$, ** $p < 0.01$, *** $p < 0.001$
Note: The number of students and teachers for this analysis was 1968 and104, respectively.

We use as an effect size indicator of the proportion of variance accounted for (R^2) derived from comparing the model with student and teacher 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 the student level is explained by student covariates—that, about two thirds of the student level variance could be due to other influences such as teacher characteristics like teacher quality (and a proportion may also be due to random variation). This gives a comparison for the amount of variance explained by teacher variance. The intraclass correlation coefficient (ICC) indicates what percent of <u>total</u> variance was due to teacher variance. High ICC values would indicate that teacher covariates could contribute 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 BEST Portfolio score.

This finding indicates that teachers who had higher portfolios scores also had greater student growth in reading comprehension, as measured by the DRP. 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). Note that, if we used the test publisher's "typical" gain over a year (between 8 and 10 units), then this proportion would be considerably smaller: 0.24. However, it is important to recall that, in the context of these urban school districts, the mean gains were found to be much smaller, than "typical," and hence, that the larger proportion is indeed a more accurate indicator.

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 of effects 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.

Conclusion

Licensure processes serve the public's interest by providing a framework for selecting qualified, competent practitioners (Kane, 2005). Put generally, the findings of this study of validity evidence for the BEST portfolio based on correlations with student achievement gains provided statistically significant but moderate evidence in support of the validity of the BEST portfolio. Our findings indicated that BEST portfolio scores do indeed allow us to distinguish among elementary teachers who were more and less successful in enhancing their students' reading achievement. HLM findings revealed that one unit change in the portfolio 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 indicated that portfolio performance was a reasonably large contributor to the total variance, but that there is still considerable variance unaccounted for. Our findings indicated further that, 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. The BEST scores contribute unique information to the prediction of students achievement gains. The fact that the

BEST scores were from just the literacy portion of the assessment, and the student assessments were also focused on literacy, but the Praxis measures covered multiple subject areas with one score, needs to be considered here too. We see this as an important result for both policy makers and researchers in the area of teacher assessment.

Our study contributes to the existing evidence base on the criterion-related validity of assessments of teaching practice by providing information about the relationships between portfolio based assessments of *beginning teachers* with a methodologically robust design. This suggests that such portfolio assessments, like those in the BEST system, could be used as an assessment of teaching practice. But so too could the observation systems studied by the MET project. In making decisions about which assessments of teaching practice to use—and how to evaluate their validity—a number of tradeoffs will need to be considered (including the feasibility of different approaches) as well as the policy uses of the assessment information. Further the use of assessment data for multiple purposes will impact decisions about the assessment instrument used and the skills and abilities measured. For example, if the assessment purpose is to serve both a summative purpose (licensure) and an "educative purpose" (mentoring) then the evidence should be structured to support both a pass/fail decision and generate evidence to provide analytic data to candidates and schools about the strength and weakness of candidate performance. Perhaps the most important question to be addressed focuses on the impact of the different approaches on the quality of teaching and learning. The goal of any evaluation system should not just be to evaluate teachers but to improve their teaching practice. That's one of the strongest arguments for including direct measures of teaching quality alongside evidence of gains in student achievement or on other evidence of student learning. It also points to the importance of considering the pedagogical value of an evaluation system—the extent to which participating in it and receiving feedback as a result supports teachers in improving their practice, and professional developers and teacher educators in supporting them.

New approaches to teacher evaluation should also take advantage of research on program practices that build teacher capacity to support greater learning such as examining the impact of induction programs. In a recent review of the literature on the impact of induction and mentoring, Ingersoll and Strong (2011) found many positive effects of induction on teacher practice and learning. Research findings from this meta-analyses of high quality induction programs showed significant positive effects on teacher satisfaction , commitment to teaching, and retention data. Further positive effects on teaching practice were also cited such as using effective questioning techniques, individualizing instruction to meet student needs, and using more effective classroom management strategies to support learning. Finally, Ingersoll and Strong's research found that students of beginning teachers' that participated in a high quality teacher induction program had higher scores on academic achievement than teachers with no induction experience. These findings suggest that induction programs that are embedded in evaluation systems that purposefully focus on building teaching capacity and are grounded in well designed evaluation systems--that ensure that evaluators are well trained, evaluation feedback is frequent, mentoring is available, and processes are in place to support struggling teachers. Putting these features in place across the lifecycle of teaching, including pre-service training, induction, and National Board certification could provide building blocks for developing a powerful human capital system that supports the collection of meaningful information about teacher effectiveness, privileges support and feedback that is well grounded in evaluation practices, and supports personnel decisions that enhance learning. This will be an important research agenda and one to which scholars of teaching evaluations are increasingly addressing.

References

Author. (2005).

Bond, Lloyd. (2001). On "Defrocking the National Board": A Reply to Podgursky. *Education Next, Fall*.

Bond, Lloyd, Smith, T., Baker, W., & Hattie, John. (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 Carolin at Greensboro.

Cavaluzzo, Linda. (2004). Is National Board Certification an effective signal of teacher quality? Alexandria, VA: The CNA Corporation.

Connecticut State Department of Education. (2006). Connecticut Education Data and Research. Retrieved January 1, 2006, from http://www.csde.state.ct.us/public/cedar/index.htm

Cunningham, George 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.

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

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. http://dx.doi.org/10.1016/0272-7757(94)00031-Z

Goldhaber, D., & Anthony, E. (2004). Can teacher quality be effectively assessed? (pp. 36). Seattle, WA: University of Washington and the Urban Institute.

Ingersoll, Richard M., & Strong, Michael. (2011). The Impact of Induction and Mentoring Programs for Beginning Teachers: A Critical Review of the Research. *Review of Educational Research, 81*(2), 201-233. http://dx.doi.org/10.3102/0034654311403323

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

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. *Educational Resources Information Center (U.S.)* [microform]. College Park, MD

Washington, DC: NPEAT University of Maryland College of Education ;

U.S. Dept. of Education Office of Educational Research and Improvement Educational Resources Information Center.

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

Measures of Effecitive Teaching (MET) Project. (2013). Ensuring Fair and Reliable Measures of Effective Teaching: Culminating Findings from the MET Project's Three-Year Study: Bill and Melinda Gates Foundation.

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

National Center for Education Statistics. (2004). Digest of education statistics, 2004. Retrieved January 2006, 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*: The National Academies Press.

National Research Council. (2008). *Assessing Accomplished Teaching: Advanced-Level Certification Programs*: The National Academies Press.

Newton, Stephen. (2010). Preservice Performance Assessment and Teacher Early Career Effectiveness: Preliminary Findings on the Performance Assessment for California Teachers. Palo Alto, CA: Stanford University

Pecheone, Ray, & Stansbury, Kendyll. (1996). Connecting teacher assessment and school reform. . *Elementary School Journal, 97*(2), 163-177. http://dx.doi.org/10.1086/461860

Podgursky, M. (2001). Defrocking the National Board: Will the imprimatur of "Board Certification" professionalize teaching? *Education Matters, Summer*(2). http://www.educationnext.org/20012/79.html

Project., Measures of Effective Teaching (MET). (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains: Bill and Melinda Gates Foundation.

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

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

Touchstone Applied Science Associates. (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.

Wenglinsky, H. (2002). How schools matter: The link between teacher classroom practices and student academic performance. *Education Policy Analysis Archives, 10*(12), 32.

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.

Youngs, Peter. (2002). State and District Policy Related to Mentoring and New Teacher Induction in Connecticut (pp. 64). New York, NY: National Commission on Teaching & America's Future.

Appendix A

OVERVIEW OF
CSDE BEST PORTFOLIO SCORING PROCESS 2006

Read the portfolio, view video tapes, and take notes.

1. READ, VIEW AND COLLECT EVIDENCE

Independent Evidence Collection

Summarize data by listing representative evidence for each performance indicator.

2. IDENTIFY PATTERNS BY PERFORMANCE INDICATOR

Identification of patterns

Identify corresponding level of performance on the feedback rubric.

SELECT RUBRIC LANGUAGE FOR FEEDBACK

Identify Rubric Language

Select appropriate level of performance for each category.

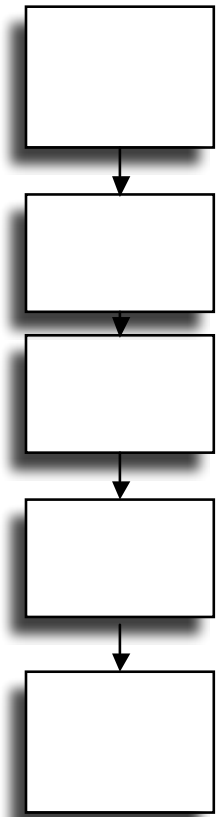
3. IDENTIFY PATTERNS AT THE CATEGORY LEVEL

Identify patterns at the category level

Review performance indicators and category patterns, identify appropriate level of performance, and assign a final score.

4. ASSIGN FINAL SCORE

Assign Holistic Score



A Table Leader reviews all steps in the written scoring process. If the Table Leader notes discrepancies across or within any documents, s/he will check with the scorer for clarification during the conference which occurs after scoring documents are handed in.

Appendix B

The framework for portfolio evaluation is organized around the following Guiding Questions which portfolio assessors use to analyze evidence from the portfolio. These questions may be used by beginning teachers to assess the quality of their own portfolios:

Category I: INSTRUCTIONAL DESIGN
How did the teacher design units in which students built understanding and applied knowledge, skills and ideas in literacy and numeracy?
I.1 Describe how the teacher used curriculum and knowledge about the students to establish expectations for learning.
I.2 Describe how the teacher focused content and learning activities to support student learning.
I.3 Describe how the teacher selected strategies and materials to support student learning.

Category II: INSTRUCTIONAL IMPLEMENTATION
How did the teacher engage students and promote their learning in literacy and numeracy?
L II.1 Describe how the teacher used reading in instruction to help students develop literacy.
L II.2 Describe how the teacher used writing in instruction to help students develop literacy.
L II.3 Describe the opportunities for students to communicate their thinking in literacy.
L II.4 Describe how the teacher differentiated instruction.
N II.1 Describe the numeracy activities used to help students problem-solve and develop numeracy.
N II.2 Describe the opportunities for students to communicate their thinking in numeracy.
N II.3 Describe how the teacher differentiated instruction.

Category III: ASSESSMENT OF LEARNING
How did the teacher communicate to students about assessment and evaluate student progress?
III.1 Describe how the teacher monitored student performance and used information about student performance in instruction.
III.2 Describe how the teacher adjusted instruction.
III.3 Describe the criteria for success.
III.4 Describe how the teacher assessed and analyzed student performance.
III.5 Describe how the teacher communicated assessment feedback to students to promote learning.

Category IV: ANALYZING TEACHING AND LEARNING
How did the teacher analyze student learning and connect it to instructional practice?
IV.1 Describe the analysis of student learning and the use of student work to support the conclusions.
IV.2 Describe how the teacher linked teaching practices to student learning.

Appendix C

Decision Guide for Holistic Evaluation of Teaching Performance

1 (• - - -)
Teacher portfolio shows little evidence that the teacher understands and implements the Professional Science Teaching Standards. Teacher focuses on the delivery of the textbook content, with little or no attention to students' learning needs and interests. Students are provided with few opportunities to experience inquiry-based science learning, and student work is focused on scientific vocabulary, with no evidence of disciplined understanding, or application of knowledge to solve problems. Teacher's reflection is weak and does not demonstrate understanding of how to facilitate the development of students' science literacy.

2 (- • - -)
Teacher portfolio shows that the teacher understands and implements the Professional Science Teaching Standards in a partial way. Teacher focuses on the delivery of the textbook content, with some attempt to address the specific needs and interests of the students. Students are provided with few opportunities to experience inquiry-based science learning, and student work is focused on the use of scientific vocabulary and basic science process skills. Teacher's reflection shows concern about students' development of science literacy, but there are no connections between analyses of students' learning and plans for future improvement of instruction.

3 (- - • -)
Teacher portfolio shows that the teacher understands and implements the Professional Science Teaching Standards. Teacher balances curricular requirements with the learning needs and interests of students and provides students with opportunities to experience inquiry-based science learning. Students' work shows use of scientific vocabulary, evidence of disciplined understanding and ability to apply scientific knowledge to solve science-related problems. Teacher's reflection shows concern about students' development of science literacy and plans for future improvement of instruction are based on analyses of students' learning.

4 (- - - •)
Teacher portfolio shows that the teacher understands and implements the Professional Science Teaching Standards in an exemplary way. Teacher balances curricular requirements with the needs and interests of students to ensure that all students learn sound science in an inquiry-based learning environment. Students' work includes multiple evidence of disciplined understanding and ability to apply scientific knowledge to solve a wide range of science-related problems. The teacher integrates instruction and assessment to facilitate student learning and evaluate their performance. Teacher's reflection shows concern about students' learning, demonstrates understanding of how to facilitate students' development of science literacy and includes thoughtful plans for future improvement of instruction.

About the Authors

Mark Wilson
University of California, Berkeley
MarkW@berkeley.edu
http://gse.berkeley.edu/people/mark-r-wilson

Mark Wilson is a professor of Education at UC, Berkeley. He received his PhD degree from the University of Chicago in 1984. His interests focus on measurement and applied statistics, and he has published just over 100 refereed articles in those areas. Recently he was elected president of the Psychometric Society, and also became a member of the US National Academy of Education, and a Fellow of the American Educational Research Association. In the past few years he has published three books: one, *Constructing measures: An item response modeling approach* (Routledge Academic), is an introduction to modern measurement; the second (with Paul De Boeck of the University of Ohio), *Explanatory item response models: A generalized linear and nonlinear approach* (Springer-Verlag), introduces an overarching framework for the statistical modeling of measurements; the third, *Towards coherence between classroom assessment and accountability* (University of Chicago Press—National Society for the Study of Education) is about the relationships between large-scale assessment and classroom-level assessment. He has also recently co-chaired a US National Research Council committee on assessment of science achievement—*Developing Assessments for the Next Generation Science Standards.*

P.J. Hallam
California Department of Education
phallam@cde.ca.gov

Dr. Hallam is an Education Program Consultant in the Professional Learning Support Division in the California Department of Education. Prior to 2011, she was a Research and Dissemination Monitor for Title II Part A, Improving Teacher Quality Grants, for the California Post-Secondary Education Commission. From 2002 to 2006, she was a post-doctorate researcher at Berkeley Evaluation Assessment and Research (BEAR) Center. Dr. Hallam's passion to learn more about the validity of large-scale educational assessments evolved as a public school teacher for fifteen years in low socio-economic communities.

Raymond Pecheone
Stanford University
pecheone@stanford.edu

Raymond Pecheone is currently a Professor of Practice in the Graduate School of Education at Stanford University. Over the course of his career, Dr. Pecheone has been a leader in high stakes educational reform through assessment, research and policy work that has shaped district and state policies in curriculum and assessment by building broad-based grassroots support for strategic new approaches to assessment and learning. Dr. Pecheone has had national impact in educational assessment through the development of nationally available assessments of teaching (edTPA) and student learning (Smarter Balanced Performance Assessment).

Pamela A. Moss
University of Michigan, Ann Arbor
pamoss@umich.edu

Pamela Moss is a Professor of Education at the University of Michigan. Her work lies at the intersections of educational assessment, philosophy of social science, and interpretive or qualitative research methods. Two edited books illustrate these intersections: *Evidence and Decision Making* (2007) illuminates the crucial roles that teachers, administrators, and other education professionals play in constructing and using evidence to make decisions that support learning. *Assessment, Equity, and Opportunity to Learn* (2008) explores the synergies and disjunctions between psychometric and sociocultural orientations to opportunity to learn and assessment. Her current research agenda focuses on validity theory in educational assessment, assessment as a social practice, and the assessment of teaching. She is a Fellow of the American Educational Research Association. She was a member of the AERA, APA, NCME committee revising the 1999 *Standards for Educational and Psychological Testing*, of the National Research Council's Committee on Assessment and Teacher Quality, and chair of AERA's Task Force on developing Standards for Reporting on Empirical Social Science Research.

education policy analysis archives

Volume 22 Number 6 February 10th, 2014 ISSN 1068-2341

arquivos analíticos de políticas educativas
conselho editorial