EVALUATION OF TEACHER PREPARATION PROGRAMS

Evaluation of Teacher Preparation Programs: A Reality Show in Kentucky

Sharon Kukla-Acevedo, Ph.D. [a]
Megan Streams, Ph.D. [b]
Eugenia F. Toma, Ph.D. [c]

[a] Corresponding Author
Central Michigan University
Department of Political Science
247 Anspach Hall
Mount Pleasant, MI 48859, USA
kukla1sa@cmich.edu
Phone: 1-989-774-3442
Fax: 1-989-774-1136

[b] Tennessee State University
College of Public Service and Urban Affairs
330 N 10th Ave., Box 139
Nashville, TN 37203, USA
mstreams@tnstate.edu

[c] University of Kentucky
Martin School of Public Policy & Administration
415 Patterson Office Tower
Lexington, KY 40506, USA
eugenia.toma@uky.edu

Abstract

Title II of the *Higher Education Act* requires states to evaluate their teacher preparation programs (TPPs). In response, many states have introduced measures to evaluate TPPs similar to the ways in which they are evaluating K-12 schools. Some states have initiated pilot projects to assess the feasibility of statewide TPP evaluations. This paper stems from the Kentucky initiative and addresses some of the methodological challenges and data limitations that surfaced when the state first considered using existing administrative data to suit this purpose. The authors outline some of these conceptual and empirical challenges in order to provide a model of learning for scholars interested in TPP evaluation and for policymakers and practitioners who are considering similar types of evaluations for their states.

KEY WORDS: Teacher preparation, evaluation, student achievement

Evaluation of Teacher Preparation Programs: A Reality Show in Kentucky

INTRODUCTION

Most people link *No Child Left Behind (NCLB)* to the intense testing of students in K-12 schools. Increased accountability of K-12 schools is the hallmark of the federal legislation. But while less known to the public, another piece of legislation may be equally important to states and education policymakers. Title II of the *Higher Education Act* (HEA) requires that states hold institutions of higher education publicly accountable for the quality of the teachers they produce. Under Title II, each state must report annually on licensure requirements, pass rates on certification assessments, state performance evaluations of teacher preparation programs, and the number of teachers in the classroom on waivers.

As a result of Title II, many states have introduced measures to evaluate teacher preparation programs (TPPs) much as they are evaluating K-12 schools. While some states such as Louisiana and Massachusetts have already begun statewide initiatives to address the Title II requirements, other states, including Kentucky, have initiated pilot projects to assess the feasibility of conducting statewide TPP evaluations using existing administrative data. This paper stems from the Kentucky initiative and addresses methodological and data issues raised by the efforts to evaluate teacher preparation programs. Administrative data from a sample of anonymous school districts within the state illustrate the challenges of this type of evaluation. The districts are both urban and rural and reveal distinct issues. This paper identifies some of the conceptual challenges of TPP evaluations in rural parts of the state and uses real-time data to illustrate issues that may have low-cost solutions, as well as those that cannot be so easily resolved. The purpose of this exercise is to serve as a model of learning for scholars interested in TPP evaluation and for policymakers and practitioners who are considering similar types of evaluations for their states. This pilot project brought a dose of reality to an abstract goal in Kentucky.

PREVIOUS WORK

Evaluation of education production has been extensive over the past thirty years. The effect of individual inputs into K-12 production have been examined to such a degree that an entire academic subdiscipline has arisen. The effectiveness of teacher salaries (Lavy, 2007; Kee & Keys, 2004; Lankford & Wyckoff, 1997; Stern, 1989), of class size (Lazear, 2001; Hoxby, 2000; Krueger, 1999; Angrist & Lavy, 1999), of school size (Andrews, Duncombe, & Yinger, 2002; Lamdin, 1995; Fowler & Walberg, 1991), of overall expenditure levels (Pritchett & Filmer, 1999; Hanushek, 1997; Hedges, Greenwald, & Laine, 1994), of the source of school revenues (Hoxby, 2001, Downes & Figlio, 1997; Downes, 1992), and increasingly of various aspects of teacher quality (Harris & Sass, 2009; Goldhaber & Anthony, 2007; Clotfelter, Ladd, & Vigdor, 2007; Betts, Zau, & Rice, 2003) on student learning have been analyzed and reviewed by academics and shared with the policy world. These studies have proven very useful as they questioned many default assumptions of those in the policy arena, such as overall per pupil spending as a source of school quality variance (Lazear, 2001; Hoxby, 2000). The studies have revealed that teacher quality appears to be one aspect of education production that both significantly affects student learning (Rivkin, Hanushek, & Kain, 2005), and is subject to policy manipulation.

While the initial studies on teachers identified teacher effects, but not which dimensions of teacher quality matter (Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004), increasingly the K-12 scholars are attempting to identify specific quantifiable elements of quality that result in increased student learning (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2006; Clotfelter et al., 2007). It should not be surprising that one element of teacher quality under examination is the training received by the teacher.

There is a vast literature dedicated to identifying what aspects of teacher preparation make a difference in student learning. Wilson, Floden, & Ferrini-Mundy (2001) provide an excellent review of the research. A major strand in this research looks at the effects of different pathways to teaching on student achievement. These studies generally compare traditional teacher education programs to alternative preparations, such as Teach for America or provisional, temporary, or emergency entry into the teaching workforce. These studies tend to find favorable effects resulting from more traditional university pathways into teaching, although the relationship is qualified by teacher experience and subject matter (Boyd, Grossman, Lankford, Loeb, and Wyckoff, 2006; Darling-Hammond, Holtzman, Gatlin, & Heilig, 2005). Student gains are generally larger (Clotfelter et al., 2007; Goldhaber & Brewer, 2000; Laczko-Kerr, 2002; Hawk, Coble, & Swanson, 1985), graduates of these programs feel more prepared (Darling-Hammond, Chung, & Frelow, 2002; Jelmberg, 1996), and they have higher classroom performance than their alternatively-certified counterparts (Good et al., 2006; Houston, Marshall, & McDavid, 2003; Hawk and Schmidt, 1989).

Research also addresses whether the amount and type of coursework a teacher completes during his or her undergraduate years later affects their students' achievement. The studies typically describe a positive, albeit nuanced, relationship between content area preparation and student test scores. Findings from Author (2009) indicate that the number of math content hours has a positive effect on student learning that does not diminish over time, while Monk (1994) suggests that the magnitude of the effect varies according to subject matter and student grade level. Goldhaber & Brewer (1997, 2000) report a positive correlation with the receipt of either a

BA or MA in math and student math achievement, but these results are not corroborated by Monk (1994) or Betts, Zau, & Rice (2003). The studies that review the effects of the amount of education courses tend to focus on outcomes other than student performance. However, both Author (2009) and Monk (1994) indicate a positive correlation between pre-service math education coursework and student achievement.

While much attention has been devoted to identifying the individual teacher characteristics and the aspects of teacher preparation that affect student learning, the broader question of whether teacher education makes a difference has only recently been pursued by researchers. The new research focus is to assess whether the education bachelors' program attended yields differences in student learning. In other words, can TPPs influence the quality of teaching or is teacher quality independent of pre-service training?

Clearly, understanding the role of TPPs is important for policies relating to both the training and licensure of teachers. There are at least three possible answers to the training question: (1) the innate characteristics of teachers determine their success in the classroom, and the training program itself has no independent effects on that success – here observed differences in the effectiveness of graduates from different TPPs would relate to differential TPP selection practices or applicant pools at those TPPs (i.e., TPP as screen, so that TPP *selection procedures* matter); (2) innate characteristics of teachers are completely dominated by the training received in the bachelor's degree TPP (i.e., TPP *instructional variation* matters); and, finally, (3) innate characteristics are important but TPPs also have an independent and significant effect on student learning (both TPP selection practices and instruction variation matter).

Policy implications differ depending on the answer. If the answer is either the extremes above, drastic changes in current training methods would be in order. For example, if the answer is (1) above, then state efforts should be directed toward identifying those characteristics in individuals and toward encouraging those persons to enter the teaching profession in a less costly way than now exists. On the other hand, (2) implies that weak entrance requirements for TPPs are not a problem, and that states' efforts should be focused on identifying spreading key best instructional practices from effective TPPs to others, or on closing ineffective TPPs altogether. Finally, if the answer is (3), both an identification of the innate characteristics of individuals who select into TPPs and an identification of specific TPP attributes that contribute to student learning are necessary.

The importance of separating the impact of individual, or innate, characteristics of the person from the effects of the educational program stem from the theoretical literature in labor economics on human capital and the signaling effects of education (Mincer, 1974; Spence, 1974; Stiglitz and Weiss, 1995). This literature has been applied in many contexts within education but the fundamental insight holds across applications. Institutions can signal quality through the selection of the best teacher candidates independently of whether they add value through the training program. Empirically, in the context of evaluating the effectiveness of TPPs, a single variable representing the institution of training with no correction for the observable and unobservable differences in individuals accepted into the training programs will lead to biased estimates on the coefficient of the TPP. This issue will be addressed in the empirical section of the paper.

Noell (2006) and Boyd, Grossman, Lankford, Loeb, and Wyckoff (2008) represent the emerging movement toward using administrative student performance data to evaluate TPPs. Noell (2006) points to evidence of differences in training programs of recent teacher graduates in Louisiana, although the author is careful to note that the project is still in the early stages of development. The state recently coordinated a substantial redesign of its teacher certification programs. The effort was backed by both state and federal funds, and one goal of the project was to implement a system that would allow the reliable quantitative assessment of the efficacy of the teachers that benefit from the redesign. The Louisiana TPP results emerge from a pilot effort that seeks to determine whether the existing administrative data can be adapted to reliably assess the effectiveness of TPPs in terms of their effect on student testing.

Similarly, a group of researchers using New York City data provide a high-resolution look at differences in teacher preparation programs in New York State. The Pathways Project received both state and federal funding to examine not only the direct effects of individual TPPs, but also the key components that these programs utilize to train first-year teachers most effectively (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2008). Using administrative and survey data, the authors find some evidence that these training differences lead to student learning differences in both Math and English Language Arts (ELA) during the teacher's first year in the classroom. They note that the programs that produce strong Math teachers also tend to produce strong ELA teachers, although some TPPs do graduate teachers that appear strong in one area only.

The current paper highlights the challenges and risks inherent in using administrative data to measure TPP efficacy by describing a pilot effort to assess the feasibility of evaluating TPPs within Kentucky. It contributes to the small, but growing literature in this area in several important ways. Public interest in assessment-based accountability will likely become increasingly inhibited by substantial state and federal budgetary pressures. TPP evaluation efforts in Louisiana and New York were made possible through additional state and federal resources; however, the Kentucky pilot project was conducted within the constraints of the concurrent budget allocation. The steps followed by Kentucky researchers will be useful information for other states attempting to perform similar evaluations, because the experience is likely to be more representative as state budget pressures increase. While Federal funding for state education departments seeking to build longitudinal data warehouses remains available on a competitive basis (National Center for Educational Statistics, 2009, Statewide Longitudinal Data Systems Grant Program, available at http://nces.ed.gov/programs/slds/), full implementation across all states will be slow, and attention needs to be given to additional fundamental concerns we identify which cannot just be solved by more data. We also describe in detail the structure and nature of Kentucky's existing administrative databases and the challenges that are inherent with this structure. This study may provide practical assistance to researchers and state policymakers with similar database structures that are interested in adapting their databases to suit this need. Along this same line, it will also provide an indication of the significant amount of information needed to provide reliable estimates of TPP value added, which is an especially difficult task in a rural state.

CONCEPTUAL MODEL AND DATA

The remainder of this paper addresses dual goals. A first goal is to assess the feasibility of introducing a TPP evaluation system in a real-world context. As such, the paper addresses the conceptual question of whether more and less successful teacher preparation programs can be identified. In other words, can typical administrative data and a typical value-added approach determine if teachers from some programs have greater classroom success than teachers from other TPP programs? Success in this case is measured strictly by student scores on state-level standardized test scores. A second goal is to use this study as a guide for other states that also are considering TPP evaluations by highlighting via empirical example the data and analytical challenges that arise in these evaluations. We identify these challenges as "reality checks" throughout the discussion which follows.

Reality Check #1: Administrative data are typically in "silos" which will make collating data needed to evaluate TPP value-added through student achievement gains difficult. Of particular interest at the outset, the State's teacher licensing agency, the Kentucky Education Professional Standards Board, recognizes 30 institutions of higher education with teacher training programs in a state with approximately 700,000 K-12 students. Among these programs, the vast majority are small, private institutions - each producing a few education graduates per year - while eight are publicly-funded institutions with substantial numbers of graduates annually. Even more to the point, four public regional comprehensive universities that began as teachers' colleges continue to dominate the market in producing teachers for the entire state. For example, Figure 1 illustrates the market for graduates of a regional institution, Eastern Kentucky University,[1] and the geographical concentration of its graduates. Figure 2 illustrates the dominance of the few institutions across the entire state, although as can be seen, individual county markets may be dominated by a small private institution. This regional segmentation of the market served by TPPs has important implications for the evaluation of their impact on student achievement, discussed below.

Like most states, Kentucky's education administrative data serve multiple purposes and the data are not collected for research priorities. Rather, each type of data is collected and coded separately by a variety of divisions within the Kentucky Department of Education (KDE) and the Education Professional Standards Board (EPSB). Figure 3 provides an overview of the different sources of administrative data. The data collected by EPSB, for example, deal only with teacher

[1] Terry Hibpshman at the Kentucky Education Professional Standards Board kindly provided these maps.

certification and licensure. These data are coded at the individual teacher level but each variable (years of experience, teacher rank, and teacher salary for example) is collected separately each year. The same applies to other data in the various administrative units. This arrangement does not appear to be unique to Kentucky and is the first “reality check” in statewide evaluations involving K-12 TPPs. States will face options of contracting out or creating research divisions within their organizational structures to collect and manage the databases necessary for these evaluations.

To illustrate the second challenge, we must consider the appropriate theoretical model that will be applied in a state’s TPP evaluation, including in this pilot project. In a typical evaluation to estimate the effects of any intervention, pre-intervention outcomes data are compared to post-intervention data, controlling for other factors that are expected to affect the outcome. The estimated coefficient on the intervention variable then signifies the magnitude of the effect of the program. Conceptually, the intervention model applied to pre-service college training is represented by the following[2]:

$$(1)\quad A_{it} = \beta_0 A_{it-1} + \beta_1 TPP_{jt} + \beta_2 Stu_{it} + \beta_3 Tch_{ijmt} + Sch_{mt} + \lambda_m + u_{it}$$

where A_i and A_{it-1} are standardized student test scores;[3] TPP_{jt} is an indicator variable designating the teacher’s preparation program; Stu_{it} is a vector of student-specific characteristics, such as student ability, race, gender, subsidized lunch eligibility, and other family socioeconomic characteristics; Tch_{ijmt} includes teacher-specific characteristics, including gender, race,

[2] The evaluation focuses on the bachelor degree programs and does not attempt to include alternative teacher pathways or master’s degree programs.
[3] An alternative specification of this equation considers the dependent variable as the change in achievement between time periods *t* and *t*-1. The gain score is a more restrictive model than the one described above. See Hanuskek (1979) for more discussion.

experience, college entrance scores, and college performance scores; and Sch_{mt} refers to time-varying characteristics of the school such as its resources, size, and student body characteristics. The subscripts denote students (i), teachers (j), schools (m) and time (t), while λ_m is a school fixed effect, and u_{it} is a randomly distributed error term. Of primary interest is the estimation of TPP, which can be interpreted as the impact of a particular teacher pre-service education on student scores, controlling for all other included factors that are expected to influence scores, including student prior scores.

Note that this model requires observations for individual students, a match of those students to their individual teachers, observations of these teacher matches with multiple students, observations of these matches over multiple teachers, and teachers from multiple TPP programs. The same students must be observed over at least two consecutive time periods so that a pre- and post- score can be calculated. If all factors that influence a student's score between periods in time are included in the empirical model, then any single estimated coefficient measures the contribution of a specific variable to the child's level of achievement in the same time period. From this perspective, it initially appears that measuring the value-added of TPP should be no different than the value added by any other policy intervention, with the exception of the additional data requirements necessary for estimation of the effects of the TPP.

In Kentucky as in most states at this time, the student-teacher matches are not available in a centralized, state location. States typically retain individual student information and individual teacher information but not in a way that enables the researcher to match the two. Rather, these matches can be made only by obtaining classroom rolls from individual schools or districts.[4]

[4] In some of the North Carolina studies, it has been assumed that the classroom teacher that administered the standardized test is the teacher of record for the student. In Kentucky, this is not the case. Teachers can request that

Kentucky has a relatively decentralized public school system with 175 school districts for its approximately 670,000 K-12 students. With the approval of the Education Professional Standards Board, three anonymous, randomly selected school districts agreed to provide their classroom rolls so that the teacher-student matches could be made. To focus the project more specifically, this evaluation observes only those students in 11th grade math courses in these three districts.[5] The 11th grade math was chosen because it coincides with the Kentucky standardized testing schedule. In particular, all 11th graders participate annually in the Kentucky Core Content Test (KCCT). The data sample for this paper uses data from students who were in 11th grade in the 2005-2006 school year.

Past standardized test results for these students are also available and are included in the analysis. Much has been written in recent years regarding the application of value-added models to assess effectiveness and researchers differ in their views on the most appropriate modeling approach (for a review, see McCaffrey, Lockwood, Koretz, & Hamilton, 2003). Selection bias may be introduced when students are not randomly placed within schools and teachers. This bias makes it difficult to separate TPPs' causal effects from the effects of pre-existing differences among classrooms with which the TPP has no relationship. One approach to mitigate these sources of bias is to use gain scores as the outcome variable of interest and/or some combination of student, teacher, and fixed effects (Boyd et al., 2008, Clotfelter et al., 2006; Harris & Sass, 2006; Author, 2009). Rothstein (2008a), however, provides some indication that a gain score may not be the most appropriate estimation strategy to model the effect of TPPs on student achievement because students' gains over the course of multiple years are dynamic and subject

---

certain students sit in their classrooms for test administration and those students are not necessarily current students of the respective teacher.

[5] Choosing a high school year proved more problematic than initially anticipated. Each of the three districts taught different math classes for 11th graders and there is a significant amount of stratification of TPP graduates across districts, as discussed later in the text. As a result, the three districts are analyzed separately.

to mean reversion. Rather, a more accurate model incorporates lagged test scores as control variables, with additional lagged scores further mitigating bias in the estimates (Rothstein, 2008b). The present study incorporates elements of Rothstein's findings by using ninth grade Math and Science scores and a 10^{th} grade Reading score as control variables in the models. The additional student test scores also serve another important function in the model. We assume these test scores control for student-specific characteristics not captured in other included variables, but which could impact the dependent variable. This makes student fixed effects unnecessary.

It is useful to examine the data in more detail. First, the three districts ranged in size from three to five high schools, with our sample of 11^{th} grade enrollments in math classes ranging from 771 to 1,699.[6] The aggregate number of math teachers in the high schools of these districts ranged from 23 to 67.[7] At the school level, the 11^{th} grade student totals range from 188 to 440. The number of math teachers whose students took the test ranged from seven to 18 per high school. Table 1 illustrates that these three districts had roughly similar math performance on the KCCT exam in the 2005-2006 school year.

Students in District 1, the smallest of the three districts, look different than those in the other two districts. As illustrated in Table 2, the district has less than one percent each of black, Asian, and Hispanic students. More males than females had scores in our sample from District

[6] These were the students eligible to be used in further analysis. These totals reflect the omission of some students from the sample for the following reasons: their math teacher taught less than five students in 2005-2006 (73 students); they changed school or district during the year (1); they had either the minimum possible or maximum possible scaled score on the exam (extreme outlying values which are identified in residual diagnostics if included in regressions) (65); or their score on the KCCT 11^{th} grade math test was missing (51). The N's of the Table 7 regressions are in turn lower than these totals due to missing data on at least one independent variable; Table 6 shows descriptive statistics for the students included in the final regression samples.
[7] Includes only teachers who taught more than five students in the sample.

1. None of these characteristics pose particular challenges for TPP evaluation. But consideration of teacher characteristic distribution brings us to the second challenge.

Reality Check #2: Dilution of TPP effects due to time since graduation is a risk, but trades off against reduction in sample size and selection effects if analysis is restricted to recent graduates only. The average years of teacher experience varies by district. Noell (2006) and others find that TPP effects are greatest during the first three years of the teacher's classroom experience. After the third year, the TPP effects are diluted by the cohort or peer effects of the school in which the teacher is hired. These peer effects are likely to be larger in schools where the teaching staff has been in place for longer periods of time. In the three districts examined, the average years of experience of high school math teachers ranges from 10 to 13 years (Table 3). Restricting the data to those teachers with experience of three years or less or, at most, to those with five years or less would cause a loss of more student observations for those schools with greater numbers of more experienced teachers. The average years of experience itself may be associated with the quality of the school, as teacher turnover is expected to be lower in schools with more amenities, such as a higher-performing student body. Thus one data choice that would make identifying TPP effects more likely – i.e., a focus on recent graduates – would entail (1) a trade-off with sample size, and (2) the risk of a school-level selection bias due to the non-random distribution of new teachers in schools from which those with more seniority tend to transfer.

Reality Check #3: (related to 2 above) If we do include all graduates (not just recent ones), the implied assumption that TPPs do not change their educational practices or selection standards over time is unrealistic. The variation in time of graduation from a teacher preparation institution presents another complication to the interpretation of estimates of TPP effects. Table 4 lists individually the three TPPs from which the largest numbers of teachers received their pre-service training and the years in which the training occurred. It groups all other Kentucky TPPs into a fourth category and then lists a fifth category for those teachers who received their undergraduate degree from a state outside of Kentucky. The data indicate that there are teachers in these districts who received degrees from TPP B as early as 1969 and as recently as 2005. If the policymakers' goal in identifying a TPP effect is to reward institutions that prepare better teachers, or to identify best curricular practices for diffusion to other institutions, one must make the assumption that the TPP effect represents some consistent educational or (less optimistically) selection practice on the part of the institution. How realistic is the assumption that a given TPP's curricular or selection approach has remained constant over such a span of years? Yet if this assumption is not met, we are again left with the implication that analysis should be restricted to only recent graduates and the same caveats that result as detailed above in Reality Check #2.

Reality Check #4: TPP grads are stratified across districts and schools: teachers graduate from TPPs and then tend to segment geographically, such that districts may get most of their teachers from one or two dominant TPPs and have none from many other TPPs. Table 5 illustrates the clustered distribution of TPP graduates across districts. Rather than grouping the teachers across all districts by institution of teacher training, Table 5 lists separately by district the number of teachers trained in the five categories of TPPs regardless of their graduation year. In District 1, for example, there are at least 21 teachers who teach one or more high school math classes and 16 of these teachers received teacher training from a Kentucky institution. Of these 16 Kentucky graduates, however, 11 were trained in the same TPP (TPP A). No high school math teachers on staff in this district in the study year were trained in TPP C. In

District 2, on the other hand, ten of 23 state-trained teachers received credentials from TPP C but no teachers were trained in the dominant training institution for District 1, TPP A. As the data illustrate, District 3 predominantly hires its teachers from yet another program, TPP B, and did not have any high school math teachers who trained in TPP A in the study year.

The data from these three districts starkly illustrate an issue that has been raised in the teacher training literature. Teachers tend to enroll in training programs near their homes and to take jobs in districts near their institutions of training (Boyd, Lankford, and Loeb, 2005). For the districts in this sample, the market segmentation is sufficiently severe that an evaluation of TPPs effectively implies a comparison of only the two most common programs, or a single dominant program vs. all others present. But because these dominant programs with sufficient numbers of graduates are not the same across the three districts, each district must be analyzed separately.[8] It is expected that enlarging the sample to include data across all Kentucky school districts would expand the set of TPPs that could be evaluated but would not mitigate the cross-district problem. Thus our final "reality check" relates to this teacher stratification issue, which appears sufficiently severe to make evaluation of TPPs in rural areas of Kentucky using this type of approach infeasible from a practical perspective.

DISTRICT LEVEL RESULTS

Recognizing the deficiencies of the data and the more fundamental "reality checks" described above, the value-added model with school fixed effects (equation 1 above) was estimated separately for each school district in the school year 2005-2006. We do not claim that this estimation solves the problems we identified above: rather, we provide an illustration of an

[8] This problem diminishes significantly in large urban areas with many teachers who trained at a variety of TPPs. Nonetheless, over 50 percent of school districts in the U.S. are rural and would be expected to share the stratification problems identified above.

attempt to estimate TPP value-added using typical available administrative data and a value-added model. Indeed, our point is that this illustration and similar approaches do *not* address those fundamental issues – and hence policymakers relying on such approaches with typical administrative data risk making unwarranted judgments about TPP efficacy.

Student test scores for end of period are their 11th grade math scores on the KCCT exam. All prior high school scores available for the students are included as controls including science and reading scores in addition to math. The student characteristics include gender, race, special abilities status, and free and reduced lunch status. The teacher's gender and experience levels are included. All school characteristics are captured in the school fixed effects[9]. The model used is therefore a version of equation (1) above based on a cross-sectional dataset of 11th grade math scores in 2005-2006.

$$(2)\quad A_{i,\,11th\ grade\ math} = \beta_0 A_{it-1} + \beta_1 TPP_{jt} + \beta_2 Stu_{it} + \beta_3 Tch_{ijmt} + \lambda_m + u_{it}$$

The subscripts denote students (i), teachers (j), schools (m) and time (t), while λ_m is a school fixed effect, and u_{it} is a randomly distributed error term.

As illustrated in Table 7, the results indicate quite clearly that a student's past scores in both reading and math (and science in the largest district) significantly influence 11th grade math scores.[10] The 10th grade reading and 9th grade math scores are significant and positively related

[9] If a panel dataset were used with multiple years of 11th grade math scores as the dependent variable, time-varying characteristics of schools could be broken out as independent variables, reflected in the Sch_{mt} term in equation (1) above. However, as we obtained one year of teacher-student matched data for this illustrative analysis, we use only the school fixed effects - which will encompass the effects for the study year of both those characteristics of schools that may change over time (ex. demographic composition) as well as any non-changing aspects (ex. physical configuration). We are not interested in differentiating the effects of particular school characteristics in this illustration.

[10] See McCaffrey, Lockwood, Koretz, & Hamilton (2003) for an excellent discussion of the issues in value-added models.

to 11th grade math performance across all districts. In fact, prior student test scores are the only variables that are consistently statistically significant in explaining student outcomes across all three districts. The significance of the school fixed effects observed in two of the districts indicates strong school-level effects, even controlling for included student and teacher characteristics.

Finally, and most important for our purposes, we consider the coefficient on the TPP variable. Recognizing that each district's equation compares only its most common TPP to all others represented by that district's high school math teachers, we see that there are no significant differences in student performance that can be attributed to the training institution. We ran an additional model that attempts to further separate the innate characteristics of the teacher from the effects of the TPP. The model includes teachers' ACT score to correct for some of the observable and unobservable differences in teachers that are accepted into the training programs. As discussed previously, an inability or failure to account for these differences in teachers, prior to selection into the TPP, will lead to biased estimates of the TPP coefficient. One complication with this model is that relatively recent hires are the only teachers in the sample for whom we have reliable ACT scores. In order to ensure a sufficient sample size, the model containing ACT scores could therefore only be run for the largest district. The District 3 student sample size was reduced from 1,137 to 370 by focusing on students of these recently graduated teachers alone. As in the previous regression, the effect of dominant TPP relative to all others remained non-significant. On the other hand, the estimated effect of teacher ACT is positive and statistically significant, even controlling for all the variables listed in Table 7.

There is more than one reason that a TPP effect may not have been identified: at the most basic level, perhaps TPP does not matter once the other included student and teacher characteristics are accounted for. We cannot make such a determination with confidence, however, due to unresolved issues in the structure of these data: if a TPP effect truly exists, the inability to identify a significant effect could be due to the fact that all teachers are included, regardless of year in which they graduated. TPPs may have varied in their selection or curricular practices over time, as discussed above. This would make our TPP signal a very noisy one. Furthermore, the approach of comparing a district's most common TPP to all others is a weak one. If the TPPs in the combined reference group have strongly divergent effects on student achievement relative to the most common TPP, it would of course be difficult to show an effect of that TPP in this design. Nonetheless, this admittedly problematic type of comparison is necessary due to the dramatic segmentation of TPPs across districts and the relatively small numbers of high school math teachers from the non-dominant TPP in each district.

CONCLUDING COMMENTS

This paper uses a unique, rich data set from anonymous school districts in Kentucky that matches teacher and school characteristics to individual 11th grade math students for the 2005-2006 school year. Despite a robust methodology and rich student-teacher level data, limitations render this study unable to determine whether some teacher training programs are better at training teachers than others in rural areas of the state, and we identify fundamental challenges which relate to this inability. To recapitulate, (1) The data requirements that would enable a successful statewide evaluation of TPP's are extensive; (2) TPP effects on teacher effectiveness are likely to attenuate with time, but restricting analysis to newer teachers will restrict sample sizes greatly and possibly introduce other selection biases due to the likely non-random distribution of seniority levels across schools with different achievement levels; (3) TPP practices may have changed over the range of graduation dates of teachers included in a value-added analysis, but analysis of only recent graduates or graduate cohorts to ameliorate this issue would, in turn, cause a further reduction in sample size and could introduce selection bias as described in (2); and (4) the most problematic issue we have highlighted – which more data alone cannot solve – is the fact that the distribution of working teachers is not independent of the location of their respective teacher training programs. At least in Kentucky, schools tend to hire disproportionately from a single TPP. Because the factors that influence the hiring may be related to those that influence student achievement and cannot be separated with these limited data, the implications of the empirical work itself are quite limited. The implications of the attempt to perform this evaluation, on the other hand, are enormous. States will have to take a close look at their data management as well as their own stratification of teacher training and hiring if they are to engage in systematic evaluations of teacher training programs. While Title II requires evaluations of TPPs, scientifically rigorous evaluations of TPP effect on student achievement using value-added methodology may be premature in most states.

Acknowledgments. The authors are indebted to the Kentucky Education Professional Standards Board for the opportunity to use administrative data that are not publicly available and to Terry Hibpshman for his role in extracting the necessary data. Jacob Fowles provided valuable assistance in manuscript preparation. We are grateful to participants of the 2008 meeting of the American Education Finance Association and the 2008 Kentucky Teacher Quality Conference for their useful comments.

Biographies

Sharon Kukla-Acevedo teaches in the Political Science Department at Central Michigan University. Her research focuses on education economics and policy in K-12 education. Current topics of interest include the evaluation of teacher effectiveness, evaluation of teacher preparation programs, teacher turnover and mobility, and improving educational outcomes for Latino/a students.

Megan Streams is an assistant professor of Public Administration in the College of Public Service and Urban Affairs at Tennessee State University. Her research interests are public finance and budgeting, with a policy interest in education; topics include the interaction of K-12 and higher education policies, impact of demographic trends on education finance, and teacher compensation and career trajectories. She has professional experience in higher education administration and budgeting.

Eugenia F. Toma is the Wendell H. Ford Professor of Public Policy in the Martin School of Public Policy & Administration at the University of Kentucky. She specializes in the economics of schooling. As a Fulbright Fellow to New Zealand in 1992, Toma worked with The Treasury and the Ministry of Education to analyze school reform. In 1998-99, she served as economics program director for the National Science Foundation (NSF) and participated in a multi-agency initiative to enhance scientific inquiry in education research. She currently is serving as PI of an NSF grant to evaluate the effectiveness of teacher professional development programs.

References

Andrews, M., Duncombe, W., & Yinger, J. (2002). Revisiting economies of size in American education: are we any closer to a consensus? *Economics of Education Review, 21*, 245-262.

Angrist, J. & Lavy, V. (1999) Using Maimonides' Rule to Estimate The Effect of Class Size on Scholastic Achievement. *Quarterly Journal of Economics 114*, 533-575.

Author. (2009).

Betts, J., Zau, A. & Rice, L. (2003). Determinants of student achievement: Evidence from San Diego. Public Policy Institute of California.

Boyd, D., Lankford, H., & Loeb, S. (2005). The Draw of Home: How Teachers' Preferences for Proximity Disadvantage Urban Schools. *Journal of Policy Analysis and Management, 24,* 1, 113-132.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2008) Teacher Preparation and Student Achievement. Working paper presented at the 2008 meeting of the American Education Finance Association, Denver, CO.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wyckoff, J. (2006). How changes in entry requirements alter the teacher workforce and affect student achievement. *Education Finance and Policy, 1,* 2, 176-216.

Clotfelter, C., Ladd, H., & Vigdor, J. (2007). How and why do teacher credentials matter for student achievement? National Bureau of Economic Research. Cambridge, MA: Working Paper #12828. Retrieved on September 21, 2007: http://www.nber.org/papers/w12828

Darling-Hammond, L., Chung, R., & Frelow, F. (2002). Variation in teacher preparation: How well do different pathways prepare teachers to teach? *Journal of Teacher Education, 53*(4), 286-302.

Darling-Hammond, L., Holtzman, J., Gatlin, S., & Heilig, J. (2005). Does teacher preparation matter? Evidence about teacher certification, Teach for America, and teacher effectiveness. Working Paper. Stanford University.

Dee, T. & Keys, B. (2004). Does merit pay reward good teachers? Evidence from a randomized experiment. *Journal of Policy Analysis and Management, 23,* 471-488.

Downes, T. (1992). Evaluating the impact of school finance reform on the provision of public education: The California case. *National Tax Journal, 45,* 405–419.

Downes, T. & Figlio, D. (1997). School finance reforms, tax limits, and student performance: Do reforms level-up or dumb down?, Tufts University, Boston.

Fowler, W. & Walberg, H. (1991). School size, characteristics, and outcomes. *Educational Evaluation and Policy Analysis*. 13, 189-202.

Goldhaber, D. & Anthony, E. (2007). Can teacher quality be effectively assessed? National board certification as a signal of effective teaching. *Review of Economics and Statistics*, *89*, 134-150.

Goldhaber, 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.

Good, T., McCaslin, M., Tsang, H., Zhang, J., Wiley, C., Rabidue Bozack, A., & Hester, W. (2006). How well do 1st year teachers teach? Does type of preparation make a difference? *Journal of Teacher Education*, *57*(4), 410-430.

Greenwald, R., Hedges, L. & Laine, R. (1996). School resources and student achievement. *Review of Educational Research*. *66*, 361-396.

Hanushek, E. (1979) Conceptual and empirical issues in the estimation of educational production functions. *Journal of Human Resources*, *14*, 351-388.

Hanushek, E. (1997). Assessing the effect of school resources on student performance: An update. *Educational Evaluation and Policy Analysis*. *19*, 2, 141-164.

Harris, D. & Sass, T. (2009). The effects of NBPTS-certified teachers on student achievement. *Journal of Policy Analysis and Management, 28*, 55-80.

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

Hawk, B., & M. Schmidt. (1989). Teacher Preparation: A Comparison of Traditional and Alternative Programs. *Journal of Teacher Education, 40*(5), 53-58.

Houston, W.R., Marshall, F., & McDavid, T. (1993). Problems of traditionally prepared and alternatively certified first-year teachers. *Education and Urban Society, 26*(1), 78-89.

Hoxby, C. (2000). The Effects of Class Size on Student Achievement: New Evidence from Population Variation. *Quarterly Journal of Economics*, *115,* 1239-1285.

_____. (2001). All School Finance Equalizations are not Created Equal. *Quarterly Journal of Economics, 115*, 1189-1231.

Jelmberg, J. (1996). College-based teacher education versus state-sponsored alternative programs. *Journal of Teacher Education, 47*(1), 60-66.

Krueger, A. (1999). Experimental Estimates of Education Production Functions. *Quarterly Journal of Economics, 114*, 497-532.

Laczko-Kerr, I., & Berliner, D.C.. (2002, September 6). 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*(37). Retrieved October 2, 2007 from http://epaa.asu.edu/epaa/v10n37/.

Lamdin, L. (1995). Testing for the effect of school size on student achievement within a school district. *Education Economics, 3*, 33-42.

Lankford, H. & Wyckoff, J. (1997). The changing structure of teacher compensation, 1970 – 1994. *Economics of Education Review, 16,* 371-384.

Lavy, V. (2007). Using Performance-Based Pay to Improve the Quality of Teachers. *The Future of Children, 17*, 1, 87-109.

Lazear, E. (2001) Educational Production. *Quarterly Journal of Economics, 116*, 777-803.

McCaffrey, D., Lockwood, J. R., Koretz, D., & Hamilton, L. (2003). *Evaluating value-added models for teacher accountability.* Santa Monica, CA: RAND.

Mincer, J. (1974). *Schooling, Experience, and Earnings,* Columbia University Press, New York.

National Center for Education Statistics. (2009). Statewide longitudinal data systems grant program: Designing, developing, implementing, & using longitudinal data systems to improve student learning. Retrieved July 30, 2009, from: http://nces.ed.gov/programs/slds/.

Noell, G. (2006). Value added assessment of teacher preparation. *Unpublished manuscript*.

Noell, G. & Burns, J. (2006). Value-added assessment of teacher preparation: An illustration of emerging technology. *Journal of Teacher Education*. 57, 1, 37-50.

Pritchett, L. & Filmer, D. (1999). What education production functions *really* show: a positive theory of education expenditures. *Economics of Education Review, 18*, 223-239

Rivkin, S., Hanushek, E. & Kain, J. (2005). Teachers, schools and academic achievement. *Econometrica*, *73*, 417-458.

Rockoff, J.E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, 94, 247-252.

Rothstein, J. (2008). Teacher quality in educational production: Tracking, decay, and student achievement. (NBER Working Paper 14442). Cambridge, MA: National Bureau of Economic Research. Retrieved November 31, 2008, from http://www.nber.org/papers/w14442.

________. (2008). Student sorting and bias in value added estimation: Selection on observables and unobservables. (NBER Working Paper 14666). Cambridge, MA: National Bureau of Economic Research. Retrieved June 27, 2009, from http://www.nber.org/papers/w14666.

Spence, A. (1974). *Market Signaling: Information Transfer in Hiring and Related Processes.* Harvard University Press, Cambridge, MA.

Stiglitz, J. & Weiss, A. (1995). "Sorting out the differences between signaling and screening theories," in M. Dempster *et. al.* (eds), *Papers in Commemoration of the Mathematical Economics Seminar at Oxford University,* Oxford University Press, Oxford.

Wilson, S., Floden, R., & Ferrini-Mundy, J. (2001). Teacher preparation research: current knowledge, gaps, and recommendations. Center for the Study of Teaching and Policy, University of Washington. Retrieved on June 27, 2009, from http://depts.washington.edu/ctpmail/PDFs/TeacherPrep-WFFM-02-2001.pdf