Evaluation of Educators and Educator Preparation Programs:

Models and Systems in Theory and Practice

WCER Working Paper No. 2014-6

November 2014

Robert Meyer, Mikhail Pyatigorsky, and Andrew Rice
Value-Added Research Center
Wisconsin Center for Education Research
University of Wisconsin–Madison
rhmeyer@wisc.edu

Wisconsin Center for Education Research
School of Education • University of Wisconsin–Madison • http://www.wcer.wisc.edu/

Meyer, R., Pyatigorsky, M., & Rice, A. (2014). *Evaluation of educators and educator preparation programs: Models and systems in theory and practice* (WCER Working Paper No. 2014-6). Retrieved from University of Wisconsin–Madison, Wisconsin Center for Education Research website: http://www.wcer.wisc.edu/publications/workingPapers/papers.php

Evaluation of Educators and Educator Preparation Programs: Models and Systems in Theory and Practice

Robert Meyer, Mikhail Pyatigorsky, and Andrew Rice

Student growth measures like value-added and student growth percentiles have become increasingly common in K-12 school and teacher accountability policy in the United States. At this time educator preparation programs are rarely evaluated by similar measures for low stakes program improvement or higher stakes accreditation processes. Growth in student achievement reflects the contributions of multiple people and multiple entities, including educator preparation programs (EPPs) via their roles in selecting educators into the profession and providing candidates with education and training. The use of teacher metrics aggregated to the EPP level to evaluate EPP performance must be viewed as a separate and distinct task from the use of value-added measures for individual teacher evaluation. Furthermore, value-added or growth measures should always be one of a suite of multiple measures used to arrive at any sort of accountability, accreditation, or program improvement decision.

The literature on EPP value-added models is fairly thin. The work that has been done focuses largely on two topics: discerning the effectiveness of teachers from different preparation programs through variously specified models and detailing model implementation as part of accountability systems. The research on value-added models and their validity as a measure of teacher or school effectiveness has grown over the past several years. The literature demonstrating the technical validity of these models has largely focused on teacher and school, not EPP, impacts. However, a growing number of articles explore value-added as a measure of EPP effectiveness.

Boyd, Grossman, Lankford, Loeb, and Wyckoff (2009) address the question of whether value-added can discern differences in the effectiveness of EPP graduates. The authors look at the effects of New York City teachers' preparation, both traditional and alternative, on their math and reading value-added results. They find meaningful differences in the average value-added across EPPs for both math and reading. Subsequent studies, however, have not been able to replicate the magnitude of these effects. A study of Washington EPPs, by Goldhaber, Liddle, and Theobald (2013) finds that EPPs account for only a small proportion of the variance in student achievement. Koedel, Parsons, Podgursky, and Ehlert (2012) come to a similar conclusion using data from Missouri.

Nonetheless, Goldhaber et al. (2013) make the argument that teacher preparation does have the potential to impact teacher effectiveness. The authors find statistically significant differences in value-added of the states' program completers. For example, they find that the spread between the value added by the average graduate of the highest and the lowest performing EPP is larger than the regression-adjusted difference between free/reduced-price lunch-eligible students and those not in the program.

A key issue in this literature is the importance of accounting for selection in model design. The selection of teachers, first into programs and then into the classrooms where they teach, can have an impact on the effectiveness ratings produced by value-added models. The various specifications explored in these papers further lead to different interpretations of the estimates produced by the models.

Mihaly, McCaffrey, Sass, and Lockwood (2012) consider model specifications specifically addressing the school/classroom selection issue. Because teachers from the same program are hired into different schools, possibly in non-random ways, growth associated with the EPP will come from different sources, such as the school environment. The authors use Florida data with a value-added model that includes school fixed effects to account for this selection. The fixed effects specification requires that graduates from different programs teach in the same schools, in order to isolate the program effects. They find that the rankings of EPPs vary greatly depending on whether school fixed effects are included.

Goldhaber et al. (2013) attempt to account for both EPP and school selection by including measures of institutional selectivity, pre-institution tests, and various school and district fixed effects: district fixed effects specifications in which program credentials are identified based on within-district differences in teachers, and school fixed effects specifications in which the differences are identified based on within-school differences among teachers. However, as the authors state:

> "[Models that include district or school fixed effects] account for time-invariant differences across schools and districts, but it is not totally clear that fixed effects models will yield unbiased estimates of mean program effects. The reason is that in a fixed-effects model, the estimates are based solely on within district or school differences in teacher effectiveness, and some of the differences between programs may be related to systematic sorting across different types of districts or schools. Imagine, for instance, that there are large differences between programs, but schools tend to employ teachers of a similar effectiveness level. In this case, a school that employs teachers that are average in effectiveness, from multiple programs, would tend to have some of the least effective teachers from the best training programs and most effective teachers from the worst training programs, and thus the within school comparison would tend to show little difference between the programs. In other words, some of the true differences between programs help explain the sorting of teachers across schools so the within school comparisons lead to a washing out of the program estimates." (pp. 32–33)

In this paper, we present a partial list of technical considerations that researchers, policy makers, and, ultimately, EPPs must address to answer such basic questions as: Which programs are more effective in recruiting and preparing effective teachers? Do traditional teacher preparation programs improve teaching quality in a measurable way?

The technical issues can be broken down into two categories: those that affect the calculation of teacher value-added, and those that affect the aggregation of teacher estimates to the EPP level.

Technical considerations discussed in this paper:

1. Teacher value-added models
    a. Pretest specifications and measurement error
    b. Student demographics
    c. Peer effects (average student demographics)
    d. Multiple years of student data
    e. Shrinkage
2. EPP value-added models
    a. Selection into jobs
    b. Match quality between teachers and jobs
    c. Time since graduation

Some of these issues have the potential to inflate the error around EPP estimates, while others can bias the estimates themselves. We believe that the most vexing problem yet to be addressed in the literature is how to account for the process through which teachers are matched up with employers (districts and schools). Under certain assumptions about the state of the teacher labor market and prevalent hiring practices, all models employed in the academic and policy spheres may be systematically underestimating the importance of teacher preparation. Section 1 presents a formal model of teacher-job sorting and demonstrates the implications for standard EPP value-added models under different assumptions about labor market conditions, hiring practices, and the extent to which districts and schools add to the quality of teaching that goes on within their classrooms. Other technical considerations listed above are discussed in section 2. Section 3 concludes.

1. Technical Considerations for EPP-Level Value-Added Models

Selection into Schools

The discussion around how best to control for selection into schools/districts has evolved but has not yet concluded. Understanding the nature of EPP-district/school networks and selection processes is important in order to disentangle the effect of district policies, school environment, and on-the-job training from the effect of teacher preparation. Failure to control for this selection

may also result in perverse incentives, reducing the motivation of EPPs and their graduates to take on challenges, such as low value-added schools or "turnaround schools." On the other hand, if schools differ in how well they can select or attract high quality candidates, then including local education agency (LEA) controls will bias EPP effect estimates toward zero, leading researchers to underestimate the importance of teacher training. Excluding LEA controls does not solve this problem; the sign of the bias is ambiguous in such a model, as we demonstrate below.

A typical teacher value-added model can be represented by the following equation:

$$Y_{ijst} = \lambda Y_{i(t-1)} + X_{it}\beta + \tau_{jt} + \varepsilon_{ijst}, \quad (1)$$

where

Y_{ijst} is the standardized test score of student *i* in year *t* attributable to teacher *j* in school *s*,

$Y_{i(t-1)}$ is that student's test score in prior period,

X_{it} is a vector of student demographic or other characteristics, such as gender, free or reduced-price lunch status, special education or English language learner classification, etc., and

τ_{jt} is the indicator variable (fixed effect) for teacher linked to student *i* in a given subject area and school year.

Dropping the time subscript, we can model teaching quality as a function of individual ability, education, and job-specific factors:

$$\tau_j = f(T_j, P_{jp}, S_{jk}), \quad (2)$$

where

T_j is that teacher's ability, distributed according to some distribution function, e.g., uniformly in the 0-1 range or normally around zero,

P_{jp} is the effect of teacher preparation program *p*, and

S_{jk} is the causal effect of working in school *k*.

In practice, these parameters can be estimated by regressing estimates of teaching quality, $\hat{\tau}_j$, from Equation 1 on EPP indicators with or without additional indicator variables (fixed effects) for the district or school (LEA). Interpretation of results, however, is complicated by the fact that "placement" of teachers in schools is not random but is rather governed by some endogenous process. Abstracting from costs (or assuming that search costs and compensation are homogenous), we can hypothesize that schools maximize expected teaching quality, $E(\tau_{jk})$, and

teachers maximize non-monetary utility of working in a given school. Under different assumptions about the true state of the world with respect to (a) how teachers select and are selected into school districts and individual schools, (b) the existence and magnitude of causal LEA effects, and (c) labor market conditions for teachers, we can show that estimates of EPP effects generated by current models may be biased. Below we describe the different states.

Case 1: No school effect, No sorting. Suppose that teachers are hired randomly, i.e., all schools are equally attractive to potential hires and schools do not observe τ_j during the hiring process (or they observe only a very noisy proxy of teaching quality).

If there is no causal school effect ($S_{jk} = 0$ for all k), i.e., a teacher's effectiveness does not depend on any school characteristics, then an unbiased estimate of P can be obtained using either overall variation in effectiveness of graduates of different programs or within-school variation. This is true regardless of whether demand exceeds supply because employment is not conditional on effectiveness.

Case 2: No school effect, Sorting. Suppose there is no causal school effect, but schools vary in their attractiveness to teachers[1] and teachers are able to signal their effectiveness to potential employees. In other words, teachers can be sorted by their effectiveness, schools can be sorted by their attractiveness, and more attractive schools will hire more effective teachers. The existence of between-school variation in teaching quality (again, driven solely by differences in effectiveness of individual teachers and not by any school-specific causal factors) means that within-school variance in effectiveness is smaller than the overall variance. If demand exceeds supply and teachers' participation constraint is satisfied (i.e., all applicants are hired by some school), the model without school fixed effects will produce an unbiased estimate of P, whereas the model with school fixed effects will be biased toward zero. If supply exceeds demand, teachers with the lowest effectiveness signals will not be hired. In this case both models will produce estimates of P that are biased toward zero;[2] the bias will remain stronger in the model with school fixed effects.[3]

Case 3: School effect, No sorting. If schools do differ in their impact on teaching quality but the sorting process is truly random, then modeling teacher preparation effects with or without controlling for the influence of schools (without LEA controls or fixed effects) should result in unbiased estimates of EPP effects. However, when applying the model to real world data (i.e., finite sample), both models may result in a bias with an unknown sign. For example, as Mihaly et al. (2012) point out, using school fixed effects assumes that schools employ teachers from multiple EPPs and that there are no systematic differences between such schools and those that employ teachers from a single EPP. The relevance of these concerns appears to vary from region to region. For example, Mihaly et al. (2012) find that more than 50% of Florida schools employ

[1] We assume that all teacher candidates share a common evaluation of what makes a school attractive.

[2] Because any such model can only produce relative estimates of program effects, attrition will lead to smaller estimates of the differences between programs and, therefore, smaller individual program coefficients.

[3] If the best teacher candidates choose to not enter the profession, the same result holds.

teachers from a single EPP, while Goldhaber et al. (2013) find a much less localized market in Washington (15% of single-EPP schools).

Case 4: School effect, Sorting. Finally, consider the case with both causal LEA effects and teacher-school sorting. A model that does not control for the LEA component of teaching quality will produce EPP estimates that are biased in the same direction as the correlation between school and program effects, corr(S,P). In other words, if "better" LEAs (ones with larger causal effects) attract more effective teachers, who are in turn more likely to be graduates of "better" programs, then the EPP estimates will be upward biased. The reverse will be true if the correlation between school and program effects is negative. This might happen if graduates of more effective programs feel better prepared to enter schools with disadvantaged student populations. Alternatively, schools that are less attractive to teachers may adopt coping strategies, such as on-the-job training, that increase their causal impact on teaching quality. Any attrition among teacher candidates, either due to low effectiveness candidates being not hired by any school or due to high effectiveness candidates switching to a different labor market, will strengthen the attenuation bias.

Just as in Case 2, adding school fixed effects does not solve this problem. Regardless of the sign of the correlation between S and P, sorting of more effective teachers into more attractive schools means that the within-school variance in teaching quality must be smaller than its sample-wide variance, meaning that estimates of P will be biased toward zero under any market conditions.

To demonstrate the potential impact of various misspecifications, we present results of a simple simulation. There are five educator preparation programs, with 100 graduates each, and 10 schools, each hiring 50 teachers.[4] EPP effects are calculated as the EPP number, 1-5, minus 3, i.e., follow a discrete uniform distribution over the -2 to 2 range with a standard deviation of 1.42. Ability of graduates of any EPP is normally distributed around 0. We allow individual ability to play a larger role than education by setting the standard deviation of the generating distribution to 3, resulting in a standard deviation in the sample of 2.89 and a significant overlap in observed teaching quality of graduates from the five EPPs (see Figure 1).

[4] To simulate excess labor supply or attrition, in some scenarios we reduce the total number of schools.

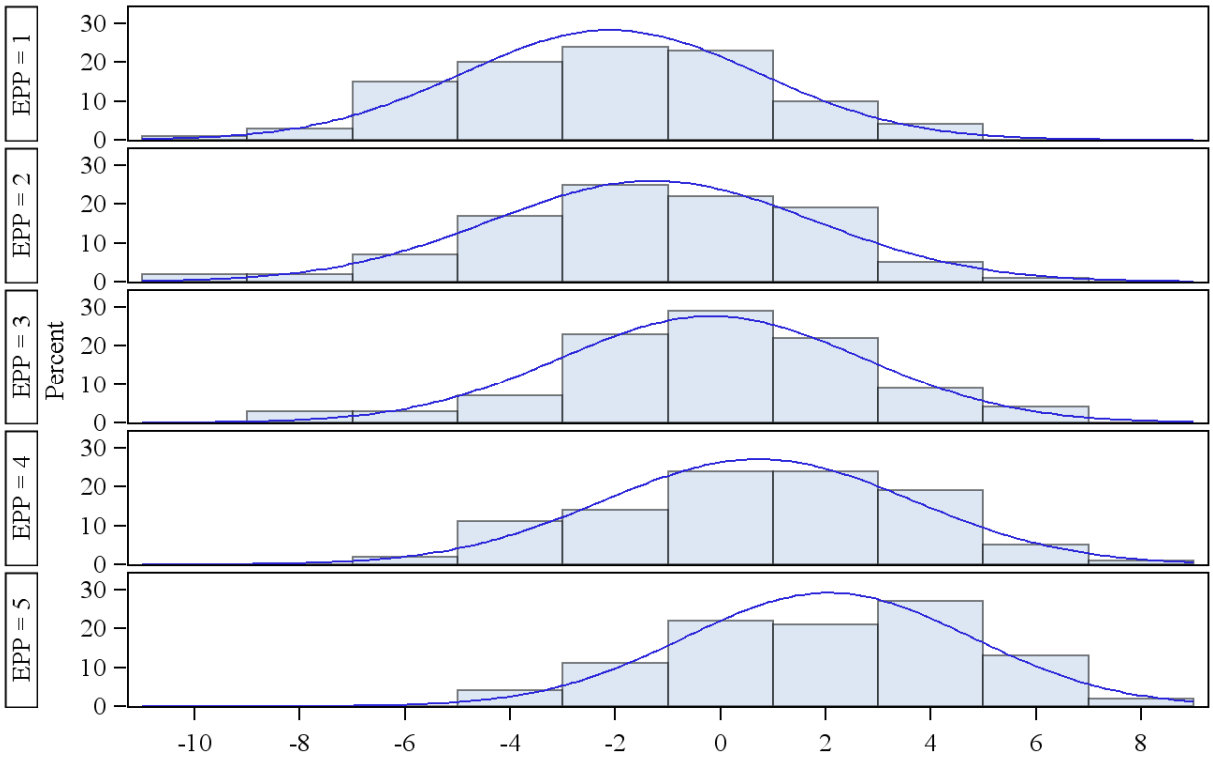

Figure 1. **Distribution of teaching quality, by educator preparation program**

School effects are set to zero or generated, similarly to EPP effects, based on the school number, 1-10, to have a mean of zero and a standard deviation of 1.44.

Finally, the sorting process is simulated in the following ways:

a. "Positive" sorting: Teacher observations are sorted based on the sum of their P and T values. Teachers with the smallest P+T value are assigned to School 1 (the school with the lowest S), etc. In this world there is no overlap in effectiveness among teachers placed in different schools (see Figure 2).

b. "Negative" sorting: Teachers are again sorted according to P+T, and those with the smallest P+T value are assigned to school 10 (the school with the *highest* S), etc.

c. "Noisy" sorting: We generate an additional "assignment error" variable, e, based on a normal distribution with a mean of zero and a standard deviation of 5. Teacher observations are then sorted and assigned based on the value of P+T+e. This does not simulate perfectly random sorting of teachers and schools, but rather a world where schools observe a noisy signal of applicants' expected effectiveness, resulting in a significant overlap in effectiveness between schools (Figure 3).

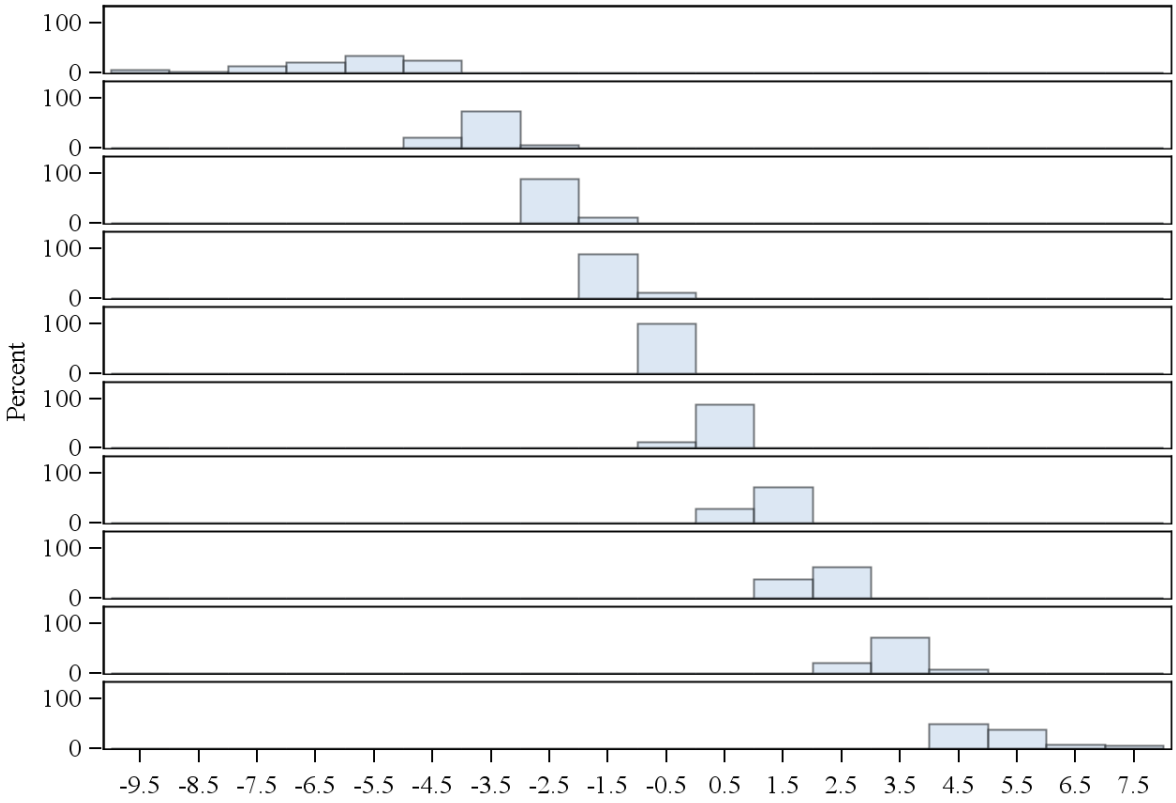

***Figure 2*. Distribution of teaching quality by school: Positive or negative sorting**

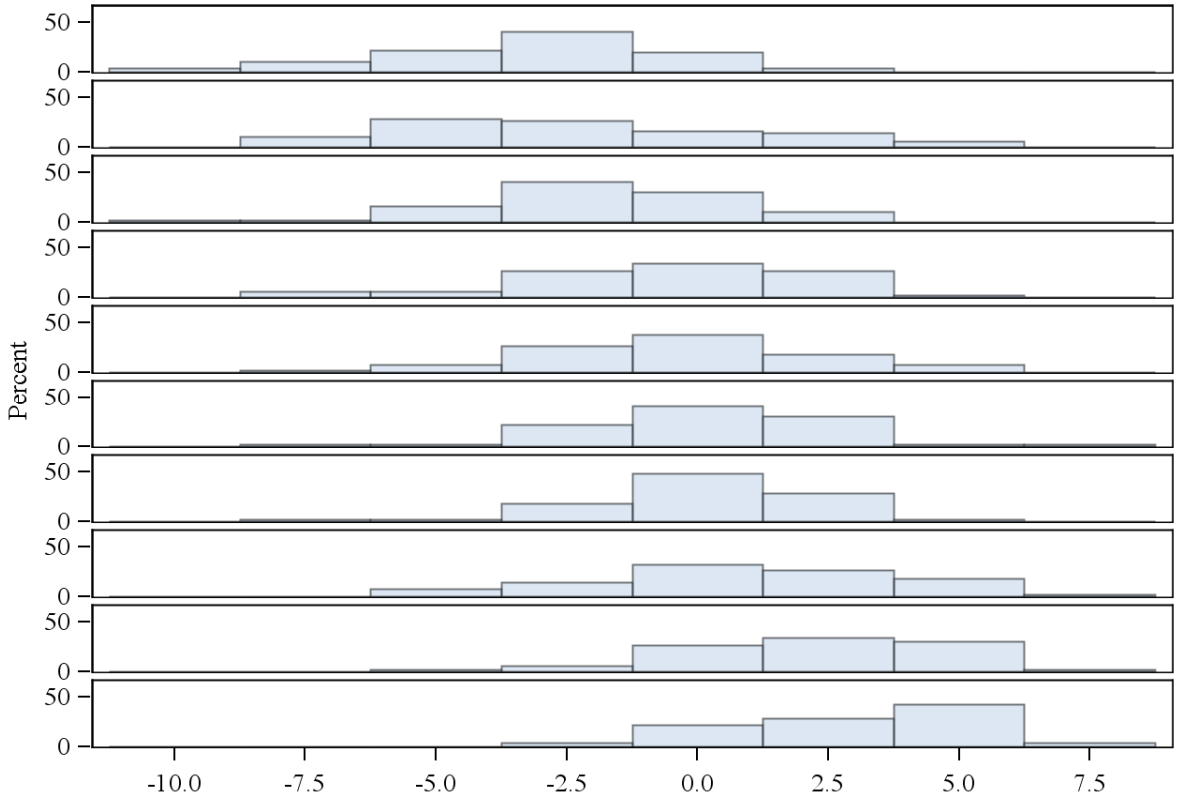

***Figure 3*. Distribution of teaching quality by school: Noisy sorting**

Table 1 summarizes the standard deviations of the different simulated variables under various scenarios. Figure 4 describes the spread (standard deviation) of EPP estimates under these scenarios (the top bar shows the standard deviation of the true EPP effects). It demonstrates two important results: (1) If teachers are able to signal their effectiveness (sorting is not "noisy"), including school fixed effects when estimating EPP effects results in a severe attenuation bias. (2) If a significant proportion of EPP graduates fails to find teaching jobs (in the relevant geographic market) or chooses to not enter the profession, the EPP model without school fixed effects may also be severely biased toward zero.

***Table 1*. Simulation assumptions – Standard deviations of model components**

	EPP Effect (P)	Teacher Ability (T)	P+T	School Effect (S)	Total Teaching Quality (τ)	Sorting Noise
No school effect, teacher sorting	1.42	2.89	3.24	0.00	3.24	0.00
School effect, "positive" sorting	1.42	2.89	3.24	1.44	4.65	0.00
School effect, "negative" sorting	1.42	2.89	3.24	1.44	1.87	0.00
No school effect, "noisy" sorting	1.42	2.89	3.24	0.00	3.24	5.04

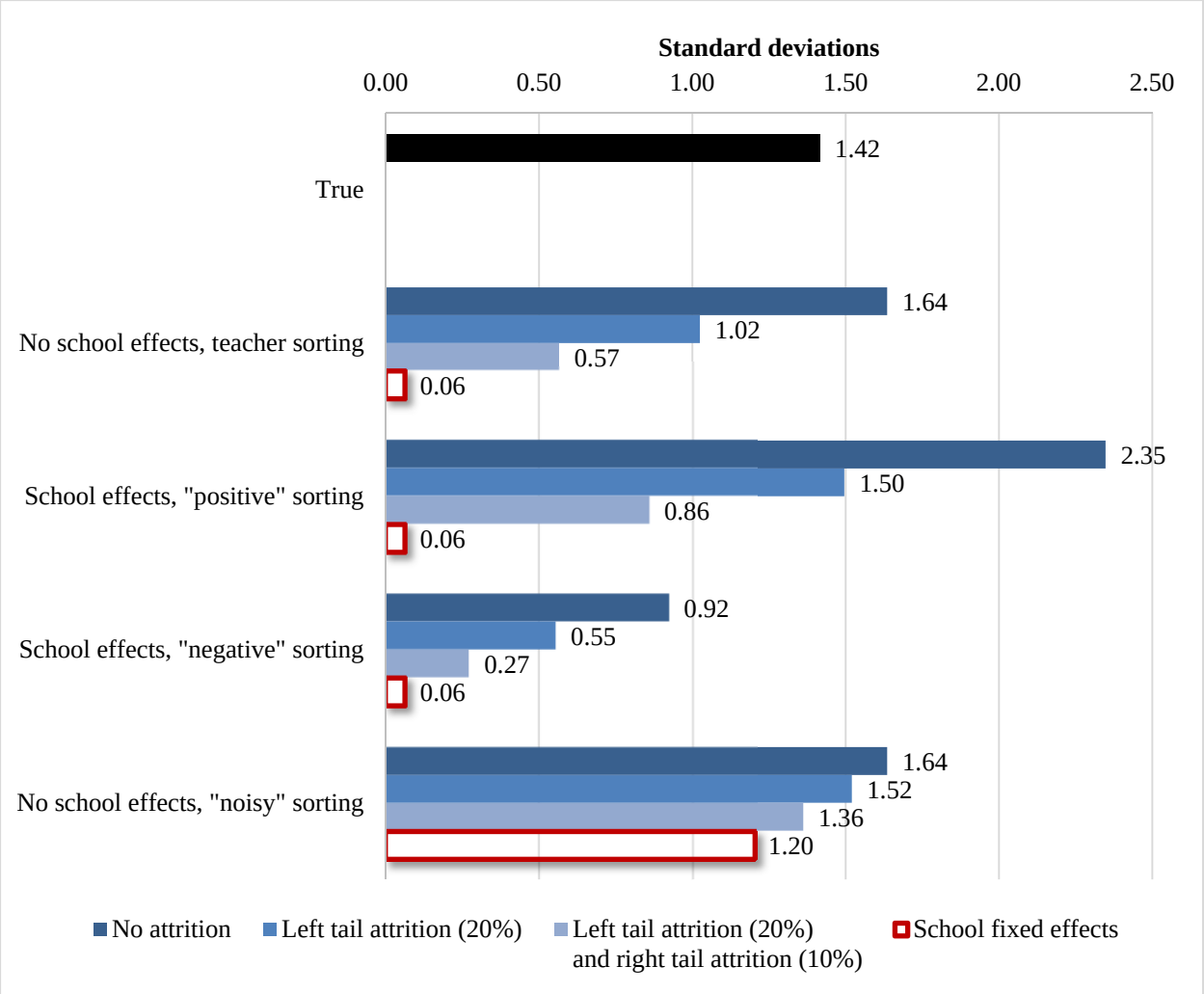

***Figure 4*. Simulation results – Standard deviations of EPP estimates**

Model Limitations

Match quality

The model of sorting presented above assumes that if causal school effects do exist, they affect overall teaching quality equally for all teachers. However, sorting can be modeled more broadly as the quality of the match between teachers and jobs. If this match quality is specific to a teacher, instead of being uniform for a given LEA or even EPP-LEA pair, then omitting information about match quality may further bias model results. For example, imagine a school responds to a shortage in a specific area, such as middle school math, by hiring some teachers with only early grades training and assigning them to middle school grades. Such teachers are poorly matched to their jobs. Is their low performance caused by the school's placement policy or by the quality of the training they received from their EPP? The answer may depend on

specifics of the distinction between "training" and certification.[5] In some cases it may be appropriate to exclude teachers who teach outside the scope of their training from analyses. For research purposes, it is advisable to model the effects of match quality instead, but the level of detail and variation needed for such an analysis may make it infeasible.

Time since graduation

Our model also abstracted from the concept of time, particularly time since graduation, which generally translates into job experience. Standard economic theories of production suggest that teacher effectiveness depends not only on training, whether pre- or in-service, but also on experience. There are several reasons why the effect of experience must be addressed, in one way or another, in EPP models:

1. Teachers' performance during their 1st year on the job is unlikely to provide a true measure of their future performance. There are also likely to be significant differences in performance during the first 1-4 years of teaching, on average.

2. Any measure of teacher performance is likely to fluctuate from year to year, suggesting a need for multi-year measures.

3. Due to existence of (a) small cohort sizes and (b) stratification in teacher labor markets, pooling analyses over multiple years appears to be a technical necessity as well.

Prior research varies widely in how authors address this topic. Boyd et al. (2009) do not include any experience controls but compare results for teachers with 1-2 years of experience to results for a larger sample of teachers with 1-5 years of experience. Mihaly et al. (2012) include broad experience categories, such as "1-2 years," "3-5 years," "6-12 years," etc. Koedel et al. (2012) also control for experience, but they do not specify exactly which functional form they utilize. Finally, Goldhaber et al. (2013) treat years of experience as a discrete variable. The authors also assume that EPP effects decay over time, at an exponential rate, and find that including the decay parameter significantly changes the magnitude of estimates for some EPPs. Allowing for decay when analyzing data pooled over multiple years is an important innovation and should be studied further.

When including teachers with more years of experience in the model, it is important to be cognizant of another selection process: attrition, i.e., the fact that we do not observe EPP graduates who exited teaching (or at least the geographic market being studied). This awareness is key because teachers who exit may be different in important ways from those who stay, e.g., they may be low performing teachers who were denied tenure (or, conversely, high performing teachers who switched professions). Goldhaber et al. (2013) investigate the possible effects of attrition and find them negligible. However, the generalizability of this finding is not yet known.

[5] For more examples, see description of the implementation process in Texas (Lincove, Osborne, Mills, and Dillon, 2013).

2. Technical Considerations for Teacher-Level Value-Added Models

Literature on teacher-level student growth and value-added models is extensive and, despite considerable controversy, numerous education agencies have integrated these models into education improvement strategies. Nonetheless, there is considerable divergence in precise applications of these models (which, one can argue, contributes to the controversy). A generalized value-added model attempts to isolate the effect of a teacher from other impacts on student growth. This section lists common features of value-added models and their impacts on teacher/classroom measures, as well as their impact on the interpretation of these measures for use in EPP accreditation or accountability.

Pretest Specifications and Measurement Error

Student growth model specifications range from assuming a known linear relationship between a pretest and posttest to a linear estimated relationship to a semiparametric nonlinear relationship. A given assumption on these issues may lead to specification error if the model is not sufficiently close to the true reality of the relationship. The following will describe the three main assumptions in the field and the possible effects relating to EPP evaluation.

Assumption 1: There is a known linear relationship between pretest and posttest (in particular, the outcome variable can be written as the difference of posttest-pretest).

If this assumption is correct, the model can proceed without bias in the estimates. If the assumption is not correct, there will be a bias in the teacher effects. In particular, if the assumed relationship is too strong, the bias will be inversely related to attainment and a teacher in a high attainment class will have artificially low measures of effect on student growth. If the relationship is assumed to be too weak, the model becomes artificially closer to an attainment model and will artificially increase measures of effectiveness for teachers that teach high attaining children (and vice versa for teachers teaching lower attaining students). Most models in the field that make this assumption tend to assume that the relationship is stronger than most estimates of the relationship.

Assumption 2: There is an unknown but estimable linear relationship between posttest and pretest.

In this case the model assumes a linear relationship but attempts to estimate the strength of this relationship using the sample data. While attractive, this practice can cause a biased result if there is no consideration of measurement error in the assessments (in particular the pretest variables as in Meyer, 1999). Not accounting for measurement error leads to a relationship that is too weak (as described above) and can shift the model results closer to an attainment model. Several methods can correct this deficiency. Teacher effects that come from a model that estimates this relationship but does not correct for assessment measurement error should be discussed in an EPP portfolio of evidence. To the best of our knowledge, all EPP papers discussed here do not attempt to correct for measurement error in assessment variables.

Assumption 3: There is an unknown but estimable general relationship between posttest and pretest.

This method does not assume the relationship between posttest and pretest is linear nor does it attempt to assume the strength of the relationship. Two particular ways of making this assumption are (a) perform a semiparametric estimation of the relationship using splines or (b) inserting a nth order polynomial into a linear regression where n may typically be up to 3 (so that there are squared and cubed terms in the model). The benefit is that if the relationship can be estimated, we can get the most accurate measures of effectiveness (especially in very high attaining and low attaining classrooms). The current drawback of these methods is that there is no generally available correction for measurement error and thus they fall prey to the underestimation of the strength of the relationship (making them closer to an attainment model).

Student Demographics

To the extent that various student demographics are correlated with the growth of test scores, it may be appropriate to control for student-level data. These variables may include free/reduced lunch status, English language learner status, special education status, ethnicity, gender, or a multitude of other variables that are collected in longitudinal data systems. The exclusion of these variables when correlated with expected test score growth causes an omitted variable bias in the teacher/classroom effect.

Inclusion of student demographics will be important for an EPP if it tends to systematically place students in very high or very low poverty areas, for example. Typically we see poverty variables limit growth of student performance. If those variables are not included in the model, an EPP in either scenario will see that teachers placed in high poverty areas have lower growth metrics simply because of model misspecification. If an EPP places a large number of teachers in high poverty areas, it could seem like the EPP trains less productive teachers.

As of this writing, two prevalent models do not usually include demographic variables: student growth percentiles as implemented in the Colorado Growth Model (many states have adopted this model), as well as the Education Value-Added Assessment System as implemented by SAS. Both of these models generally include multiple years of data in a single estimate, which mitigates bias issues in later grades; however, not enough test score data exist in the early grades to leverage for this purpose.

Classroom Average Demographic Variables

A student growth model could control for classroom or school level average demographic data to address the concern that peers may affect the growth of an individual student or group of students (a so-called peer effect).

If peer effects exist and we do not account for them in the model, they will be reflected in the performance of the teacher and the model will be biased. We sometimes call this an error of omission. Under certain conditions (i.e., when there are true causal effects of classroom average

demographics), inserting the average of classroom demographics is a good proxy for a model that controls for peer effects. If these conditions are not met (such that a non-causal relationship exists between certain average characteristics and value-added) and we insert average classroom demographics into the model, we will keep the bias and potentially mistakenly believe we are controlling for peer effects. In this case, the classroom effects cannot be interpreted causally. We call this an error of commission. Since some causal peer effects likely exist, but they may not be the dominating force in the relationship between teacher effectiveness and average classroom characteristics, the choice to include classroom average demographics is a matter of public policy, not technical method.

Many research papers in the field use classroom averages in their analysis (Kane, McCaffrey, Miller, & Staiger, 2012; Strategic Data Project, 2012; Chetty, Friedman, & Rockoff, 2012). One mechanical effect of choosing to include classroom averages is that the selection forces a zero correlation between the growth measure and any classroom average variable. Thus, if average pretest is included in the model, we will not see low growth measures and low incoming attainment linked together any more or less than low measures and high incoming attainment. Certain policies can benefit from this effect, especially where equal opportunity of success on a growth measure is required. If, however, there is generally a correlation between low attainment and low teacher effectiveness, as several studies have suggested,[6] then the model will not be able to illuminate that relationship. A recent discussion of these matters (in a state context) can be found in Ehlert, Koedel, Parsons, and Podgursky (2013).

A particular side effect of using classroom averages is that if an EPP serves particularly high achieving areas and the model used to produce teacher level growth metrics includes classroom averages, that EPP will have a collection of estimates that are biased downward (in a causal sense).

Any portfolio of evidence on student growth metrics should include information on whether classroom averages were included in the growth measure.

Multiple Years of Student Data

Some student growth models use multiple years of prior test score data in place of student demographics. It can be shown that this method does control for differences in student characteristics without using other information *given a long enough history* for the student. In most cases this means that growth models in Grades 3, 4 and 5 may not have controlled for enough student context (as the prior achievement data may not exist).

Using multiple years of student data may disadvantage certain teachers in particular contexts, and therefore could disadvantage any growth metrics tied to EPPs that trained these teachers.

[6] See, for example, Glazerman and Max (2011); Sass, Hannaway, Xu, Figlio, and Feng (2012); Isenberg et al. (2013).

There is no general way to determine the effect of these assumptions so it is recommended that an EPP accreditation system consider these factors in a portfolio of evidence.

Shrinkage

Many value-added models apply statistical shrinkage to improve the accuracy of effect estimates. Conventional empirical Bayes shrinkage estimates of productivity are appropriate for comparisons of productivity among teachers (and differences from average productivity). In particular, these methods reduce strong variance of teacher effects due to random noise caused from having low class sizes. In effect, they allow us to tame extreme measures that are due to "luck."

One important effect of applying statistical shrinkage is that teachers with small class sizes tend to be pulled into the middle of the distribution of teacher measures. This tendency makes it difficult to identify outstanding or poor performance among teachers who teach a small number of students. If a particular EPP tends to produce teachers who end up teaching very small classes, then this type of model feature may tend to squeeze that EPP's estimates toward average, making excellence more difficult to claim.

3. Conclusions

Although Boyd et al. (2009) conclude that teacher preparation is a significant component of teaching quality, more recent papers suggest that EPP effects are quite small, relative to the total variance in teachers' impact on their students' academic growth. Goldhaber et al. (2013) and Koedel et al. (2012), for example, find that the amount of variance explained by EPPs as a ratio of variance explained by individual teacher effects is on the order of 10 percent, even though the two papers disagree on what this finding says about the importance of teacher preparation. In this paper, we have shown that these estimates may be biased toward zero. We discuss a number of modeling choices and labor market conditions that can affect the magnitude and precision of results. In particular, the combination of excess supply of teachers and sorting based on effectiveness can condense the distribution of teaching quality observed among EPP graduates who actually teach. Furthermore, focusing on within-school comparison of graduates from different EPPs, i.e., putting school fixed effects into the EPP model, does not solve this problem.

School fixed effects attempt to control for observable and unobservable differences among LEAs. It is plausible that all important differences between schools *are* observable, meaning that one need only include the right observable characteristics in the model to control for the effects of selection. Mihaly et al. (2012) and Koedel et al. (2012) investigate this approach, but with a fairly narrow set of characteristics (essentially school-level averages of student demographics and achievement, and school size). The proportion of students in a school who receive free lunch or test below the proficiency level does not necessarily indicate ineffective school leadership, bad policies, or insufficient resources. For example, Harris and Sass (2011) find that having a new principal in a school correlates with lower student achievement. Further analysis with a wider set of school characteristics may be needed.

Finally, the apparent importance of labor market conditions suggests that analyses should be performed at the level of license area, instead of treating teaching as single market. The market for elementary school teachers is, usually, very different from the market for special education or high school math teachers. In fact, given sufficient sample size, these differences can be exploited to better understand the teacher-job matching process. Understanding the labor market, in turn, will help researchers state with greater certainty whether teacher preparation makes a significant impact on children's academic outcomes.

References

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis, 31*(4), 416–440.

Chetty, Raj, John N. Friedman and Jonah E. Rockoff. (2014a) "Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates." *American Economic Review* 104(9): 2593-2632.

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2013). Selecting growth measures for school and teacher evaluations: Should proportionality matter? Center for Analysis of Longitudinal Data in Education Research Working Paper 80.

Glazerman, S., & Max, J. (2011). *Do low-income students have equal access to the highest-performing teachers?* (Document No. PP11-23a). National Center on Education and the Economy Evaluation Brief. Washington, D.C.: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Goldhaber, D., Liddle, S., & Theobald, R. (2013). The gateway to the profession: Assessing teacher preparation programs based on student achievement. *Economics of Education Review*, 34, 29-44.

Harris, D. N., & Sass, T. R. (2011). Teacher training, teacher quality and student achievement. *Journal of Public Economics, 95*(7), 798–812.

Isenberg, E., Max, J. Gleason, P., Potamites, L., Santillano, R., Hock, H., & Hansen, M. (2013). *Access to effective teaching for disadvantaged students.* Washington, D.C: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Kane, Thomas J., Daniel F. McCaffrey, Trey Miller, and Douglas O. Staiger. (2013) "Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment." Seattle, WA: Bill & Melinda Gates Foundation.

Koedel, C., Parsons, E., Podgursky, M., & Ehlert, M. (2012). Teacher preparation programs and teacher quality: Are there real differences across programs? Center for Analysis of Longitudinal Data in Education Research Working Paper 79.

Lincove, J. A., Osborne, C., Mills, N., & Dillon, A.. (2012). *The politics and statistics of value-added modeling for accountability.* Paper presented at the Association for Public Policy Analysis and Management Fall Conference 2012, Baltimore, MD.

Meyer, R. H. (1999). The effects of math and math-related courses in high school. In S. E. Mayer and P. E. Peterson (Eds.), *Earning & learning: How schools matter* (pp. 169–204). Washington, D.C.: Brookings Institution Press.

Mihaly, K., McCaffrey, D., Sass, T., & Lockwood, J. R. (2012). *Where you come from or where you go? Distinguishing between school quality and the effectiveness of teacher preparation program graduates.* Georgia State University, Andrew Young School of Policy Studies Research Paper Series (12-12).

Sass, T., Hannaway, J., Xu, Z., Figlio, D., and Feng, L. (2012). Value added of teachers in high-poverty schools and lower-poverty schools. *Journal of Urban Economics, 72*, 104–122.

Strategic Data Project. (2012). *The novice teacher placement pattern.* Center for Education Policy Research, Harvard University. Retrieved from http://cepr.harvard.edu/cepr-resources/files/news-events/sdp-spi-placement-memo.pdf

Support for this research was provided by the Council for the Accreditation of Educator Preparation. Many thanks go to Emerson Elliott for comments and suggestions. Any opinions, findings, or conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the Council for the Accreditation of Educator Preparation funding agencies, WCER, or cooperating institutions. Any errors are the authors' sole responsibility.