Teacher Preparation Programs and Teacher Quality: Are There Real Differences Across Programs?

Cory Koedel
Eric Parsons
Michael Podgursky
Mark Ehlert*

July 2012

We compare teacher preparation programs in Missouri based on the effectiveness of their graduates in the classroom. The differences in effectiveness between teachers from different preparation programs are very small. In fact, virtually all of the variation in teacher effectiveness comes from within-program differences between teachers. Prior research has overstated differences in teacher performance across preparation programs for several reasons, most notably because some sampling variability in the data has been incorrectly attributed to the preparation programs.

* University of Missouri, Columbia, Department of Economics. In addition, Podgursky is a Fellow of the George W. Bush Institute at Southern Methodist University. We thank Yi Du for valuable research assistance and acknowledge support from the Center for Analysis of Longitudinal Data in Education Research (CALDER) at the American Institutes for Research. The usual disclaimers apply.

I. Introduction

There is increased national interest in holding teacher preparation programs (TPPs) accountable for how their graduates perform in the classroom. In a 2010 report from the Center for American Progress, Crowe (2010) argues that "every state's teacher preparation program accountability system should include a teacher effectiveness measure that reports the extent to which program graduates help their K-12 students to learn." This sentiment is echoed by Aldeman et al. (2011) and the United States Department of Education (2011). Numerous states have or are building the capacity to evaluate TPPs using longitudinal state data systems that link students to their teachers. In fact, all 12 winners of the federal Race to the Top (RTT) competition have committed to using student achievement outcomes for TPP evaluations, and five will use estimates of teacher impacts on student achievement for program accountability (Crowe, 2011). Some states – notably Louisiana and Tennessee – have been reporting estimates that associate TPPs with student-achievement growth for several years. The Louisiana model in particular has received considerable national attention. Paul G. Pastorek, the former state superintendent in Louisiana, shares his sentiment regarding the Louisiana model in a report released by the United States Department of Education in 2011: "I applaud the U.S. Department of Education for working to take the Louisiana-model nationwide. Teacher preparation program accountability for K-12 results is an idea whose time has come."

This paper offers a sobering view of TPP accountability of this form, at least in current application. We use a statewide longitudinal dataset from Missouri – similar to other datasets that have been used for TPP evaluations elsewhere – to examine the extent to which teachers who are prepared by different TPPs differ in effectiveness. Like other studies, we measure teacher effectiveness using value-added models (VAMs). The key result from our analysis is that teachers from different training programs differ very little, if at all, in terms of their ability to raise student achievement. That is, differences between graduates from different TPPs are small and perhaps even non-existent. Our findings suggest that TPP-of-origin is a less-useful indicator for educational administrators looking to hire effective teachers than has been implied by earlier work.

A key insight from our study is that the level of clustering of the standard errors in models of TPP effects can significantly influence the interpretation of results. Prior studies have clustered incorrectly, and in doing so have reported standard errors that are too small (e.g., Gansle et al, 2010; 2012; Noell et al., 2007 and 2008). It is important to recognize that in TPP evaluations, students who are taught by the same teacher are not independent observations regarding the effectiveness of that teacher's preparation program. Although one could argue that multiple levels of data clustering are important in models that evaluate TPPs, prior research suggests that the *most important* level is that of individual teachers.[1] After making the correct clustering adjustment, we find that most, if not all, of the variation in the estimated TPP effects in Missouri can be attributed to estimation-error variance. Prior studies have wrongly interpreted a portion of the sampling variability in the data to represent real differences in teacher effectiveness across TPPs.[2]

Our finding that true TPP effects are very small is robust to different specifications for the student-achievement model (we consider specifications that do and do not include school fixed effects) and persists despite the fact that we observe, on average, over 50 teachers per training program in our data. This is more than the average program in the Louisiana and Tennessee evaluations (Gansle et al., 2010; Tennessee Higher Education Commission, 2010). Furthermore, our findings are maintained in a subsample of large TPPs for which we observe an average of more than

[1] It has been well-established that there are large differences in effectiveness across individual teachers. See Hanushek and Rivkin (2010) for a review of the recent literature.

[2] In the Gansle et al. (2010) and Noell et al. (2007, 2008) studies, clustering occurs at the school and classroom levels, where teachers can teach in multiple classrooms. The classroom-level clustering is helpful, but will still overstate statistical power, particularly as the ratio of classrooms to teachers increases in the data. The comparison between classroom and teacher-level clustering in this context is akin to the comparison between clustering at the state and state-by-year levels in Bertrand et al. (2004). We discuss the clustering issue in more detail in Section V.

80 teachers per program. The most compelling evidence in support of the importance of clustering comes from a falsification exercise where we randomly assign teachers to TPPs (regardless of their actual TPPs of attendance). Under the random-assignment scenario with improper clustering, the models still imply non-zero differences in TPP effects. Only when we correctly cluster at the individual-teacher level do the random-assignment models show what they should – that there are no differences across the false TPPs.

The lack of true variation in TPP effects uncovered by our study is perplexing for at least two reasons. First, researchers have documented what appear to be considerable input-based differences across TPPs (Boyd et al., 2009; Levine, 2006). Given the importance of differences in effectiveness between individual teachers in determining student outcomes (Hanushek and Rivkin, 2010), and the apparent variation in program inputs across TPPs, it would seem reasonable to hypothesize that differences in how teachers are prepared across programs would translate into differences in how they perform in the classroom. We find no evidence, however, to support this hypothesis. One explanation is that the input-based differences in how teachers are trained across TPPs are not as large as they appear to be – that is, it may be that most TPPs are providing similar training.[3]

A second reason that our findings are perplexing is that our estimates also embody the effects of initial selection into the programs. So, for example, even if we take as given that differences in TPP inputs are small (in terms of producing outputs, at least), shouldn't teachers from more selective colleges perform better in the classroom? This is certainly an intuitive hypothesis, and numerous prior studies have used the selectivity of an educator's college as a proxy for quality (Clark et al., 2009; Clotfelter et al., 2006, 2007; Koedel and Betts, 2007).

[3] Regardless of the cause, our findings are consistent with TPP-of-attendance being another intervention on a long list of interventions that *do not* influence teaching effectiveness, at least at present (Glazerman et al., 2010; Harris and Sass, 2011; for a recent exception see Taylor and Tyler, 2011).

To further investigate the selection issue we compare average ACT scores for all students across Missouri public universities, and then we make the same comparison but focus only on individuals who we observe teaching in Missouri public schools. The variance of average ACT scores across universities for eventual teachers is only half as large as the variance for all students, and upon closer inspection, there is differential selection within institutions. For example, graduates from the state flagship university who end up teaching in public schools have much lower ACT scores than other students from the same university; alternatively, teachers from several other universities look similar to students who do not go into teaching. These findings offer at least a partial explanation for the recurring empirical finding that educators from more-selective institutions do not outperform educators from less-selective institutions in the classroom.

Overall, our study makes two substantive contributions to the literature on TPP evaluation. First, we advocate a technical correction for models that are used to evaluate TPPs – teacher-level clustering – and show that with improper clustering, models similar to the ones estimated here can produce misleading results. Second, after making the technical correction, we show that differences in effectiveness across graduates from different TPPs are very small. Put differently, virtually all of the variation in teacher effectiveness in the labor force occurs across teachers *within programs*. We conclude, therefore, that TPP rankings based on commonly-used value-added models are, at present, of little value to state departments of education, TPP accreditation agencies, and K-12 school administrators. If it is not made clear to K-12 administrators that the substantive differences that separate TPPs in the rankings are small, the rankings could lead to suboptimal hiring decisions because almost all of the variation in teaching effectiveness occurs within training programs.

Two qualifications to our findings are in order. First, our evaluation includes only traditional TPPs – there is likely to be additional heterogeneity across programs in analyses that also consider non-traditional TPPs (e.g., alternative-certification programs), in which case real differences in effectiveness across programs may emerge.[4] Second, our findings cannot speak to whether continued efforts to evaluate and rank TPPs based on how teachers perform in the classroom will be fruitful. For example, a recent study by The New Teacher Project (2009) suggests that the general lack of accountability within the education sector has led to complacency.[5] Indeed, our non-findings may reflect the fact that TPPs have had little incentive thus far to innovate and improve. The mere presence of annual rankings, even if they are initially uninformative, may prompt improvements in teacher preparation moving forward. Even small improvements could greatly improve students' short-term and long-term outcomes (Hanushek and Rivkin, 2010; Chetty et al., 2011; Hanushek, 2011).

II. Data

We use statewide administrative data from Missouri to evaluate recent TPP graduates from traditional, university-based programs. We began with the universe of active teachers in elementary classrooms in Missouri during the 2008-2009 school year, which is the first year for which we have linked student-teacher data. From these teachers, we identify the subset who began teaching no earlier than 2004. We then follow them for up to two additional years beyond the 2008-2009 school year, through 2010-2011, which allows us to observe up to three classrooms per elementary teacher.

We link teachers to their certification records as provided by the Department of Elementary and Secondary Education and consider all teachers who were recommended for certification by a major Missouri institution within three years of their date of first employment (consistent with the policy focus on the effectiveness of recent graduates). For the purposes of our analysis we define a

[4] Also, of course, we cannot rule out that TPP of attendance may be a more important predictor of teacher performance in other states, even among traditional TPPs. It is noteworthy, however, that the reports from Louisiana (Gansle et al., 2010; Noell et al., 2007 and 2008) and Tennessee (Tennessee Higher Education Commission, 2010) show what we interpret to be small substantive differences between teachers from different TPPs.

[5] For example, the TNTP report finds that for most teachers, areas of improvement are not identified on their annual evaluations.

"major" Missouri institution liberally – we require the institution to have produced more than 15 active teachers in our dataset. We also separately evaluate the subset of TPPs that produced more than 50 teachers in our analytic sample.[6]

Students in Missouri are first tested using the Missouri Assessment Program (MAP) exam in grade-3, which means that grade-4 teachers are the first teachers for whom we can estimate value-added to student scores. Therefore, our analysis includes all teachers in self-contained classrooms in elementary schools in grades 4, 5 and 6.[7] Our final sample includes 1,309 unique teachers who were certified from one of the 24 major preparation programs in the state. These teachers are spread across 656 elementary schools, and of those, 389 schools employ teachers from multiple programs.

In total the teachers in our sample can be linked to 61,150 students with current and lagged math test scores, and 61,039 students with current and lagged reading scores. The student-level data include basic information about race, gender, free/reduced-price lunch status, language-learner status, and mobility status (whether the student moved schools in the past year). We construct school-level aggregates for each of these variables as well, which we include in some of our models. Table 1 provides basic summary information for the data, and Table A.1 lists the teacher counts from the 24 preparation programs in our study. To maintain the anonymity of the TPPs we use generic program labels throughout our analysis.[8]

[6] A general issue in TPP evaluation is within-institution heterogeneity across programs. For example, a single institution may offer multiple certification programs across schooling levels and subjects and/or alternative certification routes. Our focus on traditional TPP programs and on teachers moving into elementary schools (a relatively homogenous output sample) reduces within-institution heterogeneity. In our primary results we simply compare all of the teachers from each program who end up in self-contained elementary classrooms in Missouri public schools. Further analysis reveals that just 1.4 percent of our main sample is identified in the certification files as obtaining an alternative certification from one of the TPPs that we evaluate. Given this low number, it is unsurprising that our findings are unaffected by our decision of whether to include these teachers in the analytic sample or not (nonetheless, results from models where alternatively-certified teachers are omitted from the analytic sample are available upon request).

[7] Our grade-6 sample is only a partial sample of grade-6 teachers in the state, as in many districts grade-6 is taught in middle schools and therefore there are no self-contained grade-6 teachers.

[8] This is at the request of the Missouri Department of Elementary and Secondary Education.

III. Empirical Strategy

Estimation of TPP Effects

We follow the empirical approach used by several other recent studies and estimate value-added models (VAMs) of the following form (Goldhaber et al., 2012; Boyd et al., 2009):

$$Y_{ijst} = Y_{ijs(t-1)}\delta_1 + X_{ijst}\delta_2 + S_{ijst}\delta_3 + T_{ijst}\delta_4 + TPP_{ijst}^{j}\theta + \gamma_s + \varepsilon_{ijst} \tag{1}$$

In (1), Y_{ijst} is a test score for student *i* taught by teacher *j* at school *s* in year *t,* standardized within grade-subject-year cell. X_{ijst} includes basic demographic and socioeconomic information for student *i*, S_{ijst} includes similar information for the school attended by student *i*, T_{ijst} includes controls for teacher experience, and TPP_{ijst}^{j} is a vector of indicator variables for the TPPs where the entry is set to one for the program from which teacher *j*, who teaches student *i,* was certified.[9] γ_s is a vector of school fixed effects; we estimate models with and without school fixed effects. We perform our analysis separately for student achievement in math and reading.[10]

Notice that the model does not include explicit controls for individual teacher effects. Of course, since our interest is in the TPP effects, we do not separate out the effects of individual teachers because the objective is to attribute teacher performance to the TPPs. However, we know from a large body of research that there are considerable differences in effectiveness across individual teachers that persist across classrooms and over time (e.g., see Hanushek and Rivkin, 2010; Goldhaber and Hansen, 2010; Goldhaber and Theobald, 2011). These differences create a clustering structure within the data. For example, if students *A* and *B* are both taught by teacher *Q*

[9] In unreported results we also verify that including additional controls for classroom characteristics does not affect our findings qualitatively.

[10] Other notable studies, Noell et al. (2007, 2008) and Gansle et al. (2010), use a multilevel model that differs mechanically from the model used here but is very similar conceptually. Goldhaber et al. (2012) also extend the general framework to account for the decay of TPP effects over time. Decay in the TPP effects is one potential explanation for why they are so small, particularly when TPP effects are estimated using data from multiple cohorts of teachers. All of the studies of which we are aware evaluate TPPs using multiple cohorts to increase teacher sample sizes. Our analysis indicates that the sample-size issue is even more important than is implied by prior studies.

who was trained at program *Z*, the two students cannot be treated as independent observations by which the effectiveness of teachers from program *Z* can be identified. Our standard errors are clustered at the individual-teacher level throughout our analysis to properly reflect the data structure.

We consider models that include several teacher characteristics, but in our primary analysis we control only for teacher experience as shown in equation (1). Research consistently shows that teacher performance improves with experience. Because we evaluate five different cohorts of entering teachers beginning in a single year (2008-2009), the experience control is important so that differences in the experience profiles of teachers across training programs are not confounded with the program impacts.[11] Goldhaber et al. (2012) show that whether the model includes controls for other observable teacher characteristics is of little practical consequence for evaluating TPPs. This is the case in our data as well (results suppressed for brevity).

We estimate the model in equation (1) with and without school characteristics, and school fixed effects, and present our findings from each specification in math and reading. We present models that compare all 24 programs and models that compare the 12 "large" programs (those that produced more than 50 teachers in our data). Whether the TPPs should be evaluated by comparing their graduates within or between schools – that is, whether the model should include school fixed effects – is unclear. A benefit of the school-fixed-effects approach is that it removes any bias owing to systematic differences across TPPs in the quality of the K-12 schools where graduates are placed. However, it also relies on comparisons between teachers at K-12 schools that house graduates from multiple preparation programs to identify the relative program impacts. That is, TPP estimates from school-fixed-effects models depend on teachers who teach in K-12 schools where teachers from *other* TPPs are also teaching. The K-12 schools that house teachers from multiple programs, and the

[11] It is uncontroversial that performance improves with experience for teachers in the early years of their careers (many studies are available – see, for example, Clotfelter et al., 2006). Recent studies by Wiswall (2010) and Papay and Kraft (2010) find that experience matters further into teachers' careers.

teachers who teach at these schools, may or may not be useful for gaining inference about the larger program effects. Mihaly et al. (2011) provide a thorough analysis of the tradeoffs involved in moving to a school-fixed-effects specification. Rather than replicate their discussion here, we simply note that these tradeoffs exist. Our key result – that there are, at most, very small differences in teacher effectiveness across TPPs – is obtained regardless of whether school fixed effects are included in our models.

Analysis of TPP Effects

After estimating several variants of the model in equation (1) for math and reading achievement, we extract the estimated TPP effects. A key policy question is this: How much do the graduates from the different TPPs differ in terms of effectiveness? This is not the same question as "what is the value-added of the training provided by different TPPs?" The latter question aims to identify the TPP effects free from the effects of initial selection into the programs. However, separating out the selection effect is unlikely to be of great interest to administrators in the field. For example, for a school district administrator, the question of *why* teachers from one TPP outperform teachers from another is not nearly as important as simply identifying the programs that, on the whole, graduate the most effective teachers. Indeed, in all of the locales where achievement-based metrics are being used to evaluate TPPs, selection effects and training "value-added" are wrapped into a single estimate. We proceed with the primary objective of estimating this combined effect, consistent with current policy practice, and return to the issue of selection into the programs below.[12]

[12] Even if we could separate out selection effects from TPP value added, it may still be desirable to evaluate TPPs based on the combined effect. For example, we may want to reward TPPs that are successful in bringing talented individuals into the teaching profession. On the other hand, in terms of developing a more effective training curriculum for teachers, understanding TPP value-added is of primary interest. Individual-level data that provide more detail about teacher experiences within TPPs, similar to the data used by Boyd et al. (2009), would be particularly valuable for learning more about what aspects of training are most important.

We produce several measures of the variability in teacher effectiveness across TPPs. The first measure is the increase in the overall (unadjusted) R-squared in the model when we add the TPP indicators. The predictive power of the TPP indicators reflects systematic differences in teacher effectiveness across graduates from different programs. If the programs do not differ, we would expect the change in R-squared to be approximately zero when the program indicators are included. We compare the change in R-squared from adding the TPP indicators to the change in R-squared from adding individual teacher indicators in their place. The change in R-squared when we add the individual teacher indicators provides a measure of the total variability in teaching effectiveness across the teachers in our data sample. The ratio of the explanatory power of the TPP indicators to the explanatory power of the individual-teacher indicators measures the share of the total variance in teacher quality that can be explained by cross-program differences.

We also estimate the variance and range of program effects. Both measures are prone to overstatement because the TPP effects are estimates; even in the absence of any real TPP effects, we would expect to estimate a non-zero variance and range of the TPP coefficients. Take the range – the value of the largest point estimate minus the smallest point estimate – as an example. Noting that each TPP coefficient is a composite of the true effect plus error, $\hat{\theta}_j = \theta_j + \lambda_j$, the range is determined partly by the estimation-error component. As the share of the total variance in the TPP effects attributable to estimation error rises, so does the overstatement of the estimated range.

Following the recent empirical literature on teacher quality, we decompose the variance in the TPP effects into two components: the true variance share and the estimation-error variance share:

$$Var(\hat{\theta}) = Var(\theta) + Var(\lambda) \tag{2}$$

In (2), $Var(\theta)$ is the true variance in the program effects, $Var(\hat{\theta})$ is the variance of the estimates from equation (1), and $Var(\lambda)$ measures the estimation-error share of the variance. We estimate $Var(\hat{\theta})$ as the raw variance of the TPP effects. We estimate $Var(\theta)$ in two ways. First, as in Aaronson et al. (2007), we estimate $Var(\lambda)$ using the average of the square of the standard errors for the program-effect estimates from equation (1). This yields a direct estimate for $Var(\theta)$ using equation (2). Second, following Koedel (2009), we use the adjusted Wald statistic from the test of the joint significance of the program effects to scale down $Var(\hat{\theta})$ and obtain $Var(\theta)$.[13]

We use the variance decompositions to approximate the share of the total variance in the estimated TPP effects that reflects actual differences in program quality. We report estimates of the adjusted standard deviation of the TPP effects using the two approaches described above. We also use the variance decompositions to adjust the range of estimates. The true range of TPP effects will always be smaller than the unadjusted range: for some unadjusted range *Z*, the expected value of the true range is $\sqrt{\frac{\text{var}(\theta)}{\text{var}(\hat{\theta})}} * Z$ (note that the true range can depend on a different pair of TPPs than the point-estimate range).[14]

[13] We also shrink the TPP effects by the ratio $\frac{\hat{\sigma}_\theta^2}{\hat{\sigma}_\theta^2 + \hat{\sigma}_{\lambda_j}^2}$, then estimate $Var(\theta)$ by estimating the variance of the shrunken estimates. $\hat{\sigma}_\theta^2$ is an estimate of the variance in the TPP effects following Aaronson et al. (2007), and $\hat{\sigma}_{\lambda_j}^2$ is the square of the standard error of the estimate for program *j*. Note that the degree of shrinkage depends on the clustering of the data through the estimated shrinkage factors. The shrunken estimates imply that the differences in TPP effects are similar to what we report below (e.g., that there are either very small or non-existent differences between TPPs).

[14] A limitation of applying the variance decompositions described above for the present application is that there are relatively few TPPs (compared to individual teachers, which is the context in which these decompositions are typically used). Nonetheless, even for the small number of TPP estimates the variance decompositions will be accurate in expectation.

IV. Results

Main Findings

Table 2 shows correlations between the TPP effects from the different models in math and reading. We estimate three different models in each subject:

> Model A: Includes the lagged student test score, student-level controls, controls for teacher experience, and the preparation program indicators
>
> Model B: Includes everything in Model A, plus school-level aggregates analogous to the student-level controls
>
> Model C: Includes everything in Model A plus school fixed effects.[15]

Table 2 shows that within subjects, the estimates from Models A and B are very similar. That is, observable differences in the K-12 schooling environments for graduates from the different TPPs introduce little bias into the estimates in Model A. Across subjects and within models, the correlation in the preparation-program effects is consistently large and positive (≈0.60), except when we include the school fixed effects in Model C (≈0.31). The estimates from Model C remain positively correlated with those from Models A and B in the same subject, but the correlations decline markedly.

One explanation for the discrepant findings between models B and C is that Model B does not capture differences in schooling environments adequately, in which case Model C would be preferred. But this explanation seems less likely given the high correlations between the estimates from Models A and B.[16] Alternatively, the differences in the estimates could reflect the failure of the

[15] We do not simultaneously include school characteristics and school fixed effects because the identifying variation by which the school-characteristic coefficients are obtained in a school-fixed-effects model comes from within-school differences, which may not be very useful over narrow time horizons. In results omitted for brevity we confirm that our main findings are qualitatively unaffected by this decision.

[16] Although the comparison between models A and B does not provide conclusive evidence with respect to the reasonableness of Model C, it is suggestive. That is, we cannot directly test for bias from unobservables in Models A or B, but the fact that the scope for bias in the estimated TPP effects from *observable* differences between schooling environments is small suggests that the role of unobserved differences may also be small (e.g., see Altonji et al., 2005).

homogeneity assumption as discussed by Mihaly et al. (2011), in which case Model B would be preferred.[17] In addition, the estimates from the school fixed effects models are noisier (the estimation-error variance roughly doubles), which causes some of the reduction in the correlations reported in the table; and the sample of teachers used to identify the TPP effects when we move to Model C changes (per the above discussion, teachers at K-12 schools where teachers from other TPPs are also observed are used to identify the TPP effects in Model C). We cannot definitively disentangle the sources of the divergence in correlations between the estimates from Models A and B, and Model C, and refer the interested reader to Mihaly et al. (2011) for further discussion. But this is largely inconsequential given our main findings, to which we now turn.[18]

Table 3 reports estimates of the variance in effectiveness across teachers from different TPPs. We split the table into two parts. The first horizontal panel evaluates the 24 "main" TPPs in Missouri (i.e., with more than 15 graduates in our data); the second horizontal panel evaluates just the "large" programs. The large program subsample includes only half of the programs but over 75 percent of the teachers in our data (Table 1). The large programs are diverse in terms of overall selectivity (based on all university entrants – see Table 4 below), and our comparisons between these programs benefit from relatively large program-level teacher sample sizes. Specifically, the average number of teachers per large program is 83.3 (Table 1).[19]

We analyze the output from each model in the same fashion throughout the table. We begin by comparing the predictive power of the TPP indicators to the predictive power of the individual teacher indicators. The first row in each panel of Table 3 shows the change in R-squared when the TPP indicators are added, and the second row shows the change in R-squared when we add

[17] Like in Mihaly et al. (2011), we also find evidence that the homogeneity assumption is violated in our data – that is, the central K-12 schools that provide the strongest connections for the TPPs are observationally different from other K-12 schools in Missouri.

[18] The correlations in Table 2 are broadly consistent with similar correlations reported by Goldhaber et al. (2012).

[19] See Appendix Table A.1 for program-by-program teacher counts.

individual-teacher indicators instead. The third row reports the ratio of the values in rows one and two. The key result is that differences in teacher performance across TPPs explain only a very small fraction of the total variance of the teacher effects. More specifically, cross-program differences explain no more than 3.2 percent of the total variance in teacher quality, and as little as one percent depending on the model. This is the first key indicator to suggest that the differences across TPPs are actually very small – almost all of the variation in teacher value-added occurs within TPPs.

The next two rows in each panel of Table 3 report the unadjusted standard deviation and range of the TPP effects from each model. The unadjusted standard deviation is calculated as the square root of the variance of the initial TPP estimates and the unadjusted range is calculated by subtracting the smallest point estimate from the largest. In each of the models where we evaluate all 24 TPPs, the unadjusted standard deviation and range are large. The unadjusted standard deviation and range are smaller, but still notable in size, when we focus on the large programs. However, neither of the unadjusted measures account for estimation error in the TPP effects.

The next two rows highlight the importance of accounting for estimation error. The variance decompositions suggest that the overwhelming majority of the variation in the raw TPP-effect estimates is the product of estimation error – that is, most of the variance is unrelated to actual differences in TPP effects. We adjust the standard deviation and range of the TPP effects in each model to account for the estimation error in the subsequent rows of the table. The label "Adjustment 1" refers to the adjustment following Aaronson et al. (2007) and the label "Adjustment 2" refers to the adjustment following Koedel (2009). Note that in our comparisons involving the large TPPs, *none of the models suggest that any of the variance across programs is real*. For the analysis of all 24 programs, the models do suggest some real differences across TPPs, but the differences are generally small, and certainly much smaller than what is suggested by the unadjusted comparisons.

Teacher Selection into TPPs

Table 3 reveals very little variation in teacher quality across teachers who attended different TPPs. The fact that the differences across TPPs are so small is surprising for at least two reasons. First is the presence of seemingly large input-based differences across teacher preparation programs (Levine, 2006; Boyd et al, 2009). However, our results are broadly consistent with other research showing teaching effectiveness can rarely be linked to *any* observable characteristic of teachers.[20] Second is that our estimates also embody differential selection into the TPPs. Even if differences in inputs across TPPs are small, one might still expect differences in teacher performance across TPPs owing to differences in selection alone.

We briefly extend our analysis to examine the selection issue in Table 4. The table uses supplementary data with information about all college graduates at the 11 public-university TPPs in Missouri that are represented in our study.[21] The first column of the table shows the average ACT score for all graduates from each university. The second column shows average ACT scores for the subset of graduates who earn an education degree.[22] The third column shows average ACT scores for the graduates who actually end up working as elementary teachers.

The bottom rows of Table 4 reveal an interesting pattern: differences in selectivity across institutions on the whole are only partly reflected within the teacher population. Put differently, teachers from colleges that are differentially selective are more similar to each other than are typical students from the same colleges. Table 4 provides at least a partial explanation for why we do not find large differences across TPPs driven by selection – the teachers from these programs are not as differentially selected as they would appear to be at first glance.

[20] See Glazerman et al. (2010) and Harris and Sass (2011); for a recent exception see Taylor and Tyler (2011).

[21] The higher-education data come from cohorts of graduates who began their college careers between the years of 1996 and 2001 and completed their degrees at one of the specified universities prior to 2009. These data do not perfectly overlap with the cohorts of teachers we evaluate but should provide a fair representation.

[22] Although not all of the teachers in our analysis are education majors, many are.

V. Clustering

Earlier in the manuscript we noted that an important difference between our analysis and several prior studies is in the level of clustering.[23] It is important to recognize that any level of clustering below the teacher level overstates independence in the data, and therefore overstates statistical precision. For example, classroom clustering assumes that classrooms taught by the same teacher represent independent observations. But teachers have persistent effects across classrooms and because of this, students who are taught by the same teacher – even across classrooms and/or years – cannot be viewed as independent observations regarding the effectiveness of that teacher's assigned TPP. The extent to which classroom clustering will overstate statistical precision depends on the persistence of teacher effects across classrooms, and the ratio of classrooms to teachers in the data. As teachers are observed with more and more classrooms, the overstatement of statistical power by models that cluster at the classroom level will increase.[24]

To illustrate the importance of clustering correctly, in Table 5 we replicate our analysis from Table 3 in math and communication arts without clustering the data, as well as with classroom-level clustering. For brevity we report findings for Model B only. Table 5 implies large differences in TPP effects in the full models, and moderately-sized differences even among the large programs. Unsurprisingly, the model without any clustering suggests that the differences across teachers from

[23] Gansle et al. (2010, 2012) and Noell et al. (2007, 2008) cluster at the classroom level. Boyd et al. (2009) do not indicate a level of clustering in their study. In correspondence with the authors we were told that clustering occurs at the teacher level, although their standard errors are much smaller than the standard errors that we report here, despite their using a smaller estimation sample and models that include school fixed effects (which typically result in larger standard errors; see Table 3). We do not have an explanation for their findings. Finally, note that a conventional wisdom is that clustering should occur "at the level of the intervention," but clustering at the TPP level will result in unreliable standard errors in TPP evaluations because the number of data clusters is less than the number of parameters to be estimated (the parameters to be estimated are the coefficients for each TPP *plus* the coefficients for the other control variables in the model). Additionally, we question the rationale for clustering at the TPP level. There is little reason to expect that two students taught by two different teachers (perhaps at different schools) belong in the same cluster, particularly when one conditions on the TPP effects directly.

[24] Again, the comparison between classroom and teacher-level clustering in this context is akin to the comparison between clustering at the state and state-by-year levels in Bertrand et al. (2004). In both intermediate cases (classroom-clustering in the present application, or state-by-year clustering in the Bertrand et al. study), the clustering structure assumes too many independent observations, leading to standard errors that are too small.

different TPPs are larger. Recall from Section II that we observe up to three classrooms per teacher given the structure of our data, and on average we observe 2.4 classrooms per teacher in the analytic sample. Note that in analyses that span longer time horizons, or where teachers teach multiple classes per year (e.g., middle or high school), the overstatement of statistical power from classroom clustering will be larger than what is shown in Table 5.

Is the issue with TPP evaluation one of sample size? That is, if we could observe more teachers from each program could we statistically identify differences in TPP effects? Trivially, of course, it will be easier to detect small differences in TPP effects with larger sample sizes, but our sample sizes are not small for this type of analysis. Again, for the large programs in Missouri we observe an average of more than 80 teachers per program. This number is larger than what is reported in the 2010 reports from Louisiana (Gansle et al., 2010) and Tennessee (Tennessee Higher Education Commission, 2010). Our sample sizes appear to be very reasonable, and even large, in terms of what can be expected from these types of evaluations.

But suppose we could increase our TPP-level sample sizes. How much would it help? From the R-squared analysis presented in Table 3, a reason to suspect that it may not help very much is that there is considerable variability in teaching effectiveness *within* programs, but very little across. The large within-program variability, combined with the small cross-program differences, suggests that the data requirements that would be required to statistically distinguish the TPP effects would be substantial.

We provide some insight into this issue in Table 6. In the first column of the table we display the TPP effects for the large programs in Missouri, in random order to maintain program anonymity, as estimated by model B for mathematics.[25] It is reasonable to interpret some of the

[25] The presentation in Table 6 is designed to maintain TPP anonymity at the request of the Missouri Department of Elementary and Secondary Education. We include the omitted program in the output – we assign it a coefficient

point estimates as large – indeed, if the unadjusted range of 0.116 represented the actual difference between the best and worst programs it would be quite meaningful. But per the preceding analysis, and as shown in the bottom rows of the table, the unadjusted range is entirely the product of estimation-error variance. To investigate further we perform the following empirical falsification test: we take all of the teachers from the 12 large programs, pull them out of their actual programs, randomly assign each teacher to a new program, and then re-estimate the model. In the random-assignment scenario teachers do not actually attend the programs to which they are assigned unless by coincidence. Put differently, the TPP assignments in column 2 are false. Therefore, we should estimate TPP effects of zero.

However, the point estimates in column 2 of Table 6 are similar in magnitude to the point estimates in column 1, as shown by the standard deviation and range (although the ordering in column 2 is unimportant given the random assignment). That is, we obtain similarly sized "TPP effects" even if we assign teachers to TPPs at random. This illustrates the point that we can obtain results similar in magnitude to what we report using teachers' actual assignments if we use false assignments.

In column 3 we look to see how much we can expect to reduce our standard errors by increasing within-program teacher sample sizes. To do this, we pull in teachers from all 24 TPPs and group them at random into 12 false TPPs of equal size. The increase in our teacher sample going from column 2 to column 3 is more than 30 percent – to an average of 109.1 teachers per TPP. Notice, however, that our standard errors are still large in column 3 despite the increase in sample size. One might hope that with large sample sizes – over 100 teachers per program – we could estimate "precise" zeros in the random-assignment scenario, but this is not the case.

estimate of zero and a standard error that is the minimum of all other standard errors in the model (although the omitted program is not the largest). We are purposefully somewhat vague here to protect the anonymity of the programs.

Finally, in column 4 we reproduce the same point estimates as in column 3, but we do not cluster our standard errors. Consistent with our earlier findings, the unclustered model suggests that some of the differences between the false-TPP effects are highly significant. This is the most compelling evidence that we can provide regarding the danger of drawing inference from the models beyond what is facilitated by the data. None of the false "TPP effects" in column 4 could possibly be real – but some imply large differences in performance across teachers from these randomly assigned groups. In results omitted for brevity, we verify that the findings in column 4 can be replicated consistently by repeating the random-assignment procedure.[26]

We conclude that given the actual data conditions under which TPP evaluations are likely to occur, and even using sample sizes on the high end of what can be expected, TPP effects cannot be statistically distinguished. The problem is the combination of too much variability in teaching effectiveness within programs and too little variability across programs.[27]

VI. Discussion and Conclusion

We evaluate TPPs in Missouri using value-added models similar to those used in previous research studies (Boyd et al., 2009; Goldhaber et al., 2012) and at least two ongoing statewide evaluations (in Louisiana and Tennessee). The work we perform here is along the lines of what has been encouraged by the United States Department of Education (2011) and scholars from the Center for American Progress (Crowe, 2010) and Education Sector (Aldeman et al., 2011), among

[26] Again, the in-between case is classroom-level clustering. We also estimate models where we cluster at the classroom level with random assignment, and unsurprisingly, our standard errors are in between what we report in columns (3) and (4) of Table 6. The estimates are on the margin of joint statistical significance (i.e., some random draws imply statistically significant differences across TPPs, other don't), which is why we don't show results from any particular draw in the table. But note that as the classroom-to-teacher ratio increases beyond the value of 2.4, which is the ratio in our data, the likelihood of identifying false TPP effects with classroom clustering will increase.

[27] Gansle et al. (2010) report that their TPP estimates are "generally consistent" (p. 21) with those from their previous reports. This seemingly contradicts our finding that the TPP estimates are virtually entirely comprised of estimation error. There are two possible explanations. First, differences across TPPs could actually be larger in Louisiana – however, the point estimates provided by Gansle et al. do not strongly support this hypothesis. A more likely explanation is that the Louisiana reports use an overlapping sample of teachers from year to year.

others. Moreover, all twelve Race-to-the-Top winners have committed to using achievement data for public disclosure of the effectiveness of TPP graduates, and five winners have committed to using teacher effects on student achievement for program accountability (Crowe, 2011). A key finding from our study, and one that we feel has not been properly highlighted in previous studies and reports, is that the measureable differences in effectiveness across teachers from different preparation programs are very small. The overwhelming majority of the variation in teacher quality occurs within programs.[28] We encourage policymakers to think carefully about our findings as achievement-based evaluation systems, and associated accountability consequences, are being developed for TPPs.

Our study also adds to the body of evidence showing that it is difficult to identify which teachers will be most effective based on pre-entry characteristics, including TPP of attendance. That said, work by Boyd et al. (2009) suggests that variation within programs in terms of preparation experiences may be large. For example, Boyd et al. find that better oversight of student teaching for prospective teachers is positively associated with success in the classroom later on. The current research literature is too thin to fully understand how differences in within-program experiences affect teaching performance, but it would not be unreasonable, for example, to expect that within-program variability in the quality of training exceeds across-program variability (in the student-teaching oversight example, this would be the case if the attentiveness of prospective teachers' mentors is unrelated to TPP of attendance).

Another area of inquiry that we investigated in some detail is with regard to initial selection into TPPs – we show that differential selection across teachers from different TPPs, based on ACT

[28] The only major state-level evaluation of which we are aware that is unlikely to overstate statistical power is the Tennessee evaluation (Tennessee Higher Education Commission, 2010). The Tennessee evaluation aggregates t-statistics from estimated individual-teacher effects to compare TPPs, so the TPP-level sample size is the number of teachers. The results from the Tennessee evaluation are substantively similar to our own.

scores, is much smaller than what is implied by comparing the umbrella institutions that house the TPPs. That is, teachers who are trained at different TPPs are more similar to each other than the typical, non-teaching students who attend the universities where the TPPs are located. This may help to explain our substantive finding that there are not large differences in the effectiveness of graduates from different TPPs – the selection dimension does not appear to be as important among would-be teachers as would be expected based on institution-level differences in student selection across all fields of study.

We conclude by noting that our findings need not be interpreted to suggest that formal, outcome-based evaluations of TPPs should be abandoned. In fact, the lack of variability in TPP effects could partly reflect a general lack of innovation at TPPs, which is facilitated by the absence of a formal evaluation mechanism. The mere presence of an evaluation system, even if it is not immediately fruitful, may induce improvements in teacher preparation that could improve students' short-term and long-term outcomes in meaningful ways (Chetty et al., 2011; Hanushek, 2011; Hanushek and Rivkin, 2010). Still, we caution researchers and policymakers against overstating the present differences in TPP effects as statewide rankings become increasingly available. If administrators do not understand how small the differences in TPP effects really are, they could make poor hiring decisions by overweighting TPP rankings in their decisions.

References

Aaronson, Daniel, Lisa Barrow and William Sander. 2007. Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics* 25(1), 95-135.

Aldeman, Chad, Kevin Carey, Erin Dillon, Ben Miller and Elena Silva. 2011. A Measured Approach to Improving Teacher Preparation. Education Sector Policy Brief.

Altonji, Joseph G., Todd E. Elder and Chistopher R. Taber. 2005. Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools. *Journal of Political Economy* 113(1): 151-184.

Bertrand, Marriane, Esther Duflo and Sendhil Mulllainathan. 2004. How Much Should We Trust Differences-in-Differences Estimates? *Quarterly Journal of Economics* 119 (1), 249 – 275.

Boyd, Donald, Pam Grossman, Hamilton Lankford, Susanna Loeb and Jim Wyckoff. 2009. Teacher Preparation and Student Achievement. *Educational Evaluation and Policy Analysis* 31(4): 416-440.

Chetty, Raj, John N. Friedman and Jonah E. Rockoff. 2011. The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood. National Bureau of Economic Research Working Paper No. 17699.

Clark, Damon, Paco Martorell and Jonah Rockoff. 2009. School Principals and School Performance. CALDER Working Paper No. 38.

Clotfelter, Charles T., Helen F. Ladd and Jacob L. Vigdor. 2006. Teacher-Student Matching and the Assessment of Teacher Effectiveness. *Journal of Human Resources* 41(4): 778-820.

--. 2007. Teacher Credentials and Student Achievement: Longitudinal Analysis with Student Fixed Effects. *Economics of Education Review* 26(6): 673-682.

Crowe, Edward. 2011. Getting Better at Teacher Preparation and State Accountability: Strategies, Innovations and Challenges Under the Federal Race to the Top Program. Policy Report. Center for American Progress.

--. Measuring What Matters: A Stronger Accountability Model for Teacher Education. Policy Report. Center for American Progress.

Gansle, Kristin A., Noell, George H., R. Maria Knox, Michael J. Schafer. 2010. Value Added Assessment of Teacher Preparation in Lousiana: 2005-2006 to 2008-2009. Unpublished report.

Glazerman, Steven, Eric Isenberg, Sarah Dolfin, Martha Bleeker, Amy Johnson and Mary Grider. 2010. Impact of Comprehensive Teacher Induction: Final Results from a Randomized Controlled Study. Policy Report. Mathematica Policy Research.

Goldhaber, Dan and Michael Hansen. 2010. Using Performance on the Job to Inform Teacher Tenure Decisions. *American Economic Review (P&P)* 100(2), 250-255.

Goldhaber, Dan and Roddy Theobald. 2011. Managing the Teacher Workforce in Austere Times: The Implications of Teacher Layoffs. CEDR Working Paper.

Goldhaber, Dan, Stephanie Liddle and Roddy Theobald. 2012. The Gateway to the Profession: Assessing Teacher Preparation Programs Based on Student Achievement. CEDR Working Paper.

Hanushek, Eric A. 2011. The Economic Value of Higher Teacher Quality. *Economics of Education Review* 30(3), 466-479.

Hanushek, Eric A. and Steven G. Rivkin. 2010. Generalizations about Using Value-Added Measures of Teacher Quality. *American Economic Review (P&P)* 100(2), 267-271.

Harris, Douglas N. and Tim R. Sass. 2011. Teacher Training, Teacher Quality and Student Achievement. *Journal of Public Economics* 95(7-8): 798-812.

Koedel, Cory. 2009. An Empirical Analysis of Teacher Spillover Effects in Secondary School. *Economics of Education Review* 28(6): 682 – 692.

Koedel, Cory and Julian R. Betts. Re-Examining the Role of Teacher Quality in the Educational Production Function. University of Missouri Working Paper No. 07-08.

Levine, Arthur. 2006. Educating School Teachers. Policy Report. The Education Schools Project.

Mihaly, Kata, Daniel McCaffrey, Tim R. Sass and J.R. Lockwood. 2012. Where You Come From or Where You Go? Distinguishing Between School Quality and the Effectiveness of Teacher Preparation Program Graduates. CALDER Working Paper No. 63.

Noell, George H., Bethany A. Porter, R. Maria Patt. 2007. Value Added Assessment of Teacher Preparation in Lousiana: 2004-2006. Unpublished report.

Noell, George H., Bethany A. Porter, R. Maria Patt, Amanda Dahir. 2008. Value Added Assessment of Teacher Preparation in Lousiana: 2004-2005 to 2006-2007. Unpublished report.

Papay, John P. and Mathew A. Kraft. 2010. Do Teachers Continue to Improve with Experience? Evidence of Long-Term Career Growth in the Teacher Labor Market. Unpublished Manuscript, Harvard University.

Staiger, Douglas, Jonah Rockoff (2010). Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives* 24(3), 97-118.

Taylor, Eric S. and John H. Tyler. The Effect of Evaluation on Performance: Evidence from Longitudinal Student Achievement Data of Mid-career Teachers. NBER Working Paper No. 16877.

Tennessee Higher Education Commission. 2010.Report Card on the Effectiveness of Teacher Training Programs. Report.

The New Teacher Project. 2009. The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness. TNTP Policy Report.

United States Department of Education. 2011. Our Future, Our Teachers: The Obama Administration's Plan for Teacher Education Reform and Improvement.

Wiswall, Mathew. 2011. The Dynamics of Teacher Quality. Unpublished manuscript. New York University.

Table 1. Data Details.

Primary Dataset	
Preparation Programs Evaluated	24
New teachers in grades 4, 5 and 6 who were verified to receive a certification and/or degree from a valid institution within three years of start date*	1309
Maximum number of teachers from a single preparation program	143
Minimum number of teachers from a single preparation program	16
Number of schools where teachers are observed teaching	656
Number of schools where teachers from more than one preparation program are observed teaching	389
Number of students with math test score records who could be linked to teachers	61150
Number of students with com test score records who could be linked to teachers	61039
Average number of classrooms per teacher	2.38
Programs Producing 50 or More New Teachers†	
Preparation Programs Evaluated	12
New teachers in grades 4, 5 and 6 who were verified to receive a certification and/or degree from a valid institution within three years of start date*	1000
Maximum number of teachers from a single preparation program	143
Minimum number of teachers from a single preparation program	50
Number of schools where teachers are observed teaching	555
Number of schools where teachers from more than one preparation program are observed teaching	363
Average number of classrooms per teacher	2.38

* Note that we only include teachers who were in self-contained classrooms in these grades (e.g., elementary schools). Many grade-6 teachers teach in middle schools in Missouri.

† Program 12 was also included in this group, although it only had 49 new teachers represented in the sample (see Appendix Table A.1).

Table 2. Correlation Matrix for Preparation-Program Effects Estimated from Different Models.

24 Major Preparation Programs	Math A	Math B	Math C	Com-Arts A	Com-Arts B	Com-Arts C
Math A	1.00					
Math B	0.98	1.00				
Math C	0.53	0.52	1.00			
Com-Arts A	0.57	0.52	0.22	1.00		
Com-Arts B	0.61	0.62	0.22	0.96	1.00	
Com-Arts C	0.06	0.05	0.31	0.25	0.27	1.00
Student Covariates	X	X	X	X	X	X
School Covariates		X			X	
School Fixed Effects			X			X

Notes: Models A, B and C are as described in the text.

Table 3. Variation in Preparation Program Effects.

	Math			Communication Arts		
	Model A	Model B	Model C	Model A	Model B	Model C
24 Major Preparation Programs						
ΔR-Squared from Adding TPP Indicators	0.001	0.001	0.001	0.001	0.001	0.000
ΔR-Squared from Adding Teacher-Level Indicators	0.045	0.045	0.019	0.027	0.026	0.012
Ratio	0.024	0.027	0.032	0.030	0.027	0.009
Unadjusted Standard Deviation of TPP Effects	0.039	0.041	0.058	0.034	0.035	0.028
Unadjusted Range of TPP Effects	0.154	0.162	0.205	0.161	0.161	0.125
Estimation-Error Variance Share (Adjustment 1)	0.966	0.852	0.752	0.750	0.691	1.00
Estimation-Error Variance Share (Adjustment 2)	0.923	0.831	0.858	0.758	0.703	1.00
Adjusted Standard Deviation (Adjustment 1)	0.007	0.016	0.029	0.017	0.019	0
Adjusted Standard Deviation (Adjustment 2)	0.011	0.017	0.022	0.017	0.019	0
Adjusted Range (Adjustment 1)	0.028	0.062	0.102	0.079	0.089	0
Adjusted Range (Adjustment 2)	0.042	0.067	0.077	0.079	0.088	0
Programs Producing 50 or More New Teachers						
ΔR-Squared from Adding TPP Indicators	0.001	0.001	0.000	0.000	0.000	0.000
ΔR-Squared from Adding Teacher-Level Indicators	0.047	0.047	0.017	0.027	0.026	0.011
Ratio	0.019	0.019	0.012	0.015	0.011	0.009
Unadjusted Standard Deviation of TPP Effects	0.031	0.031	0.024	0.022	0.019	0.022
Unadjusted Range of TPP Effects	0.125	0.116	0.090	0.087	0.071	0.072
Estimation-Error Variance Share (Adjustment 1)	1.00	1.00	1.00	1.00	1.00	1.00
Estimation-Error Variance Share (Adjustment 2)	1.00	1.00	1.00	1.00	1.00	1.00
Adjusted Standard Deviation (Adjustment 1)	0	0	0	0	0	0
Adjusted Standard Deviation (Adjustment 2)	0	0	0	0	0	0
Adjusted Range (Adjustment 1)	0	0	0	0	0	0
Adjusted Range (Adjustment 2)	0	0	0	0	0	0

Notes: Estimation Error Adjustment 1 refers to the first procedure outlined in the text that follows Aaronson, Barrow and Sander (2007). Adjustment 2 refers to the second procedure that follows Koedel (2009). Standard errors are clustered at the teacher level in all models. In cases where the error-variance adjustment implies that the total-variance share that can be explained by error variance exceeds 1.00, a value of 1.00 is reported.

Table 4. Average ACT Scores by University for 11 Public Universities Included in Our Evaluation.

		Average ACT Scores	
	All Graduates	Graduates with Education Major	Observed Elem Teachers
Univ of Missouri-Columbia*	26.3	25.7	24.2
Univ of Missouri-Kansas City	25.3	23.8	22.8
Missouri State Univ*	24.7	24.1	21.7
Missouri Southern State Univ*	23.9	24.0	21.1
Univ of Missouri-St. Louis*	23.7	23.0	22.0
Southeast Missouri State Univ*	22.9	23.2	21.9
Northwest Missouri State Univ*	22.6	22.7	22.8
Univ of Central Missouri*	22.6	22.5	20.8
Missouri Western State Univ*	22.5	23.3	21.0
Lincoln University	21.4	22.0	21.0
Harris-Stowe State University	19.3	19.8	18.8
Variance of ACT Scores Across Universities	3.68	2.17	1.93
Range of ACT Scores Across Universities	7.0	5.9	5.4

* Indicates that the program is one of the twelve large programs in the state.

Notes: The calculations in columns (1) and (2) are based on graduates from the listed universities who entered the public system between 1996 and 2001 and graduated prior to 2009. The calculations in column (3) are based on data from the teachers whom we evaluate in our study. The population standard deviation in ACT scores, nationally, is approximately 4.5. The estimated TPP effects are purposefully omitted from this table.

Table 5. Replication of Model B in Math and Communication Arts. Standard Errors are not Clustered.

Clustering	Math Model B None	Math Model B Classroom	Comm Arts Model B None	Comm Arts Model B Classroom
24 Major Preparation Programs				
ΔR-Squared from Adding TPP Indicators	0.001	0.001	0.001	0.001
ΔR-Squared from Adding Teacher-Level Indicators	0.045	0.045	0.026	0.026
Ratio	0.027	0.027	0.027	0.027
Unadjusted Standard Deviation of TPP Effects	0.041	0.041	0.035	0.035
Unadjusted Range of TPP Effects	0.162	0.162	0.161	0.161
Estimation-Error Variance Share (Adjustment 1)	0.176	0.574	0.237	0.511
Estimation-Error Variance Share (Adjustment 2)	0.167	0.569	0.244	0.510
Adjusted Standard Deviation (Adjustment 1)	0.037	0.027	0.031	0.024
Adjusted Standard Deviation (Adjustment 2)	0.037	0.027	0.030	0.024
Adjusted Range (Adjustment 1)	0.147	0.106	0.141	0.113
Adjusted Range (Adjustment 2)	0.148	0.106	0.140	0.113
Programs Producing 50 or More New Teachers				
ΔR-Squared from Adding TPP Indicators	0.001	0.001	0.000	0.000
ΔR-Squared from Adding Teacher-Level Indicators	0.047	0.047	0.026	0.026
Ratio	0.019	0.019	0.011	0.011
Unadjusted Standard Deviation of TPP Effects	0.031	0.031	0.019	0.019
Unadjusted Range of TPP Effects	0.116	0.116	0.071	0.071
Estimation-Error Variance Share (Adjustment 1)	0.197	0.709	0.502	1.00
Estimation-Error Variance Share (Adjustment 2)	0.180	0.664	0.461	0.978
Adjusted Standard Deviation (Adjustment 1)	0.027	0.017	0.014	0
Adjusted Standard Deviation (Adjustment 2)	0.028	0.018	0.014	0.003
Adjusted Range of TPP Effects (Adjustment 1)	0.104	0.063	0.050	0
Adjusted Range of TPP Effects (Adjustment 2)	0.105	0.067	0.052	0.011

Notes: See notes to Table 3.

Table 6. Investigation of Program-Effect Magnitudes and Standard Errors. Programs with at least 50 New Teachers. Model B. Math.

	Actual Estimates	Estimates from Random Assignment Scenario	Estimates from Random-Assignment Scenario with Increased Sample Sizes	Estimates from Random-Assignment Scenario with Increased Sample Sizes *Unclustered*
Program 1	0.047	0.030	0.041	0.041
	(0.028)	(0.028)	(0.030)	(0.012)
Program 2	-0.010	-0.002	0.029	0.029
	(0.030)	(0.033)	(0.028)	(0.012)
Program 3	-0.006	0.029	0.038	0.038
	(0.028)	(0.032)	(0.028)	(0.012)
Program 4	-0.067	0.069	0.022	0.022
	(0.032)	(0.038)	(0.030)	(0.012)
Program 5	0.022	-0.011	-0.001	-0.001
	(0.026)	(0.027)	(0.027)	(0.012)
Program 6	0.004	0.000	0.057	0.057
	(0.028)	(0.029)	(0.031)	(0.012)
Program 7	0.023	0.005	0.055	0.055
	(0.038)	(0.033)	(0.029)	(0.012)
Program 8	-0.016	0.034	0.051	0.051
	(0.043)	(0.042)	(0.030)	(0.012)
Program 9	0.005	-0.013	0.047	0.047
	(0.027)	(0.025)	(0.029)	(0.012)
Program 10	0.011	0.022	0.044	0.044
	(0.033)	(0.032)	(0.028)	(0.011)
Program 11	0.049	0.002	0.033	0.033
	(0.033)	(0.029)	(0.029)	(0.012)
Program 12	0.000	0.000	0.000	0.000
	(0.026)	(0.025)	(0.027)	(0.011)
Avg Number of Teachers per TPP	83.3	83.3	109.1	109.1
Teacher Dispersion	As in Data	As in Data	Uniform Across Programs	Uniform Across Programs
Unadjusted St Dev	0.031	0.024	0.020	0.020
Unadjusted Range	0.116	0.083	0.059	0.059
Adjusted St Dev (Adj 1)	0	0	0	0.015
Adjusted Range (Adj 1)	0	0	0	0.046

Notes: The program labels are randomly assigned within the programs that produced at least 50 new teachers in our sample – they cannot be linked to the counts in Appendix Table A.1. In the last three columns, the actual TPP designations are not relevant – teachers are assigned at random to artificial TPPs. In the last two columns, the artificial TPPs are constructed to be of equal size; column 2 maintains the relative sizes of the programs from column 1. In the final two columns, we randomly re-allocate all teachers in our entire sample to an artificial TPP. The error adjustments for the standard deviation and range of TPP effects are based on adjustment 1 from Table 3.

Appendix A
Supplementary Tables

Appendix Table A.1. Teacher Counts for Teacher Preparation Programs in Missouri that Produced More than 15 Teachers in our Final Analytic Sample.

Program	New Elementary Teacher Count (from the Analytic Data Sample)
Program 1	143
Program 2	120
Program 3	118
Program 4	111
Program 5	106
Program 6	76
Program 7	61
Program 8	57
Program 9	53
Program 10	53
Program 11	53
Program 12*	49*
Program 13	39
Program 14	39
Program 15	34
Program 16	30
Program 17	29
Program 18	26
Program 19	24
Program 20	20
Program 21	19
Program 22	17
Program 23	16
Program 24	16

* Note that we include program 12 in our “large” program sample, although our findings are not qualitatively sensitive to dropping it from this group.