Published in 2016 by *Economics of Education Review* 53:31-45.

Teacher Quality Differences Between Teacher Preparation Programs: How Big? How Reliable? Which Programs Are Different?

Paul T. von Hippel
University of Texas at Austin

Laura Bellows
Duke University

Cynthia Osborne
University of Texas at Austin

Jane Arnold Lincove
Tulane University

Nick Mills
formerly University of Texas at Austin

Acknowledgments: This research began under a contract with the Texas Education Agency (TEA). The conclusions are the authors', not TEA's.

Abstract

Sixteen US states have begun to hold teacher preparation programs (TPPs) accountable for teacher quality, as estimated by teacher value-added to student test scores. Yet it is not easy to identify TPPs whose teachers are substantially better or worse than average. The true differences between TPPs are small; the estimated differences are not very reliable; and when many TPPs are compared, multiple comparisons increase the danger of misclassifying ordinary TPPs as good or bad. Using large and diverse data from Texas, we evaluate statistical methods for estimating teacher quality differences between TPPs. The most convincing estimates come from a value-added model where confidence intervals are widened by the Bonferroni correction and by the inclusion of teacher random effects (or teacher clustering in large TPPs). Using these confidence intervals, it is rarely possible to tell which TPPs, if any, are better or worse than average. The potential benefits of TPP accountability may be too small to balance the risk that a proliferation of noisy TPP estimates will encourage arbitrary and ineffective policy actions.

1 Introduction

After years of holding individual teachers accountable for their effects on student learning, policy leaders have raised their sights to the programs in which teachers are prepared. While governments have long played a role in approving and funding teacher preparation programs (TPPs), sixteen states have begun to practice a more rigorous form of TPP accountability, which has higher stakes and is more focused on results.

The purpose of the new TPP accountability is to "close failing [TPPs], strengthen promising programs, and expand excellent programs" (Levine, 2006; cf. US Department of Education, 2011). In Texas, for example, the State Board of Educator Certification is now authorized to warn a TPP, to put a TPP on probation, to assign a TPP to intervention, or to revoke a TPP's accreditation. The Board is also required to post estimates of TPP quality on the internet, providing "consumer information" that, like college rankings, can guide aspiring teachers in deciding which TPP will train them, and guide school administrators in deciding between job candidates from different TPPs (Texas State Legislature, 2009).

To assess TPP quality, the new accountability systems "focus on student achievement as the primary measure of success" (Levine, 2006). A "good" TPP is defined as one whose teachers raise student test scores and graduation rates more than teachers from other TPPs. Defining TPP quality in terms of student outcomes is a sharp break with older systems that defined quality in terms of TPP inputs, resources, or processes. For example, as of 2006, states approved and accredited TPPs primarily on the basis of their coursework and student teaching requirements. About a third of states required faculty to hold a doctorate, and about a third also required prospective teachers to pass an admission or graduation test and to exceed a threshold grade point average (GPA) (Levine, 2006, Table 14). Under the new accountability, a TPP's training methods and the grades or test scores of its trainees are

secondary issues. The primary question is whether the TPP is turning out teachers who raise student achievement.

While a policy of holding TPPs accountable for the effects of their teachers on student achievement may seem promising, several conditions must be met for it to work in practice. The first condition is that teachers from different TPPs must differ substantially in their effectiveness. The average difference between teachers from good and bad TPPs must be large enough that a decision to expand a good TPP or close a bad one would have a meaningful effect on student achievement. This is not a given. Although individual teachers vary substantially in effectiveness, it may be that little of the variation in teacher effectiveness lies between TPPs.

A second condition for effective accountability is that it must be possible to estimate the differences between TPPs reliably—i.e., without too much estimation error or noise. Noise adds to the variation in TPP estimates and makes the differences between TPPs appear larger than they truly are. In addition, noise makes it hard to tell which TPPs are better or worse. If estimated TPP differences are very noisy, then a TPP's position at the top or bottom of the rankings may have more to do with random estimation error than with true quality, and policies based on TPP rankings will be arbitrary and ineffective.

A third condition for effective TPP accountability is that we must be able to identify with confidence the individual TPPs that are better or worse than average. Singling out good and bad TPPs is not a trivial matter. It is possible to accept the global hypothesis that TPPs differ in their effects, and yet remain uncertain about which individual TPPs are better or worse. Noise in the estimated TPP differences is just one problem. Another problem is *multiple tests* (Hsu, 1996). We can test each TPP estimate for significance, but if we conduct multiple hypothesis tests at a significance level of .05, then purely by chance we would expect to conclude that 5 of the nearly 100 TPPs in Texas differ

significantly from the average—even if all were truly identical. To avoid basing policy decisions on random chance, it is necessary to correct for multiple tests. This correction will inevitably reduce the number of TPPs that appear to be different.

In short, the potential of a TPP accountability system hinges on the three questions in our title:

1. How big are the teacher quality differences between TPPs?
2. How reliably can those differences be estimated?
3. How confidently can we single out individual TPPs as different?

The answers to these questions have changed over time. Early TPP evaluations in New York City and Louisiana suggested that there were large teacher quality differences between TPPs, and that those differences could be reliably detected despite noise in the estimates (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2009; Gansle, Noell, & Burns, 2012a). But more recent TPP evaluations in Missouri and Washington state suggested that true teacher quality differences between TPPs were quite small (Goldhaber, Liddle, & Theobald, 2013; Koedel, Parsons, Podgursky, & Ehlert, 2015)—in fact indistinguishable from zero in some analyses (Koedel, Parsons, et al., 2015). The Missouri evaluation estimated that most of the variation between TPP estimates consisted of noise rather than true differences in teacher quality (Koedel, Parsons, et al., 2015). No TPP evaluation has considered the problem of multiple tests.

While it is possible that the differences between TPPs are larger in some states than in others, it is also possible that the divergent conclusions of past TPP evaluations were due in part to methodological decisions. Past research has highlighted the sensitivity of TPP estimates to decisions about which covariates to include, whether to include school fixed effects (FEs), and how to cluster standard errors (SEs) (Koedel, Parsons, et al., 2015; Lincove, Osborne, Dillon, & Mills, 2014; Mihaly, McCaffrey, Sass, & Lockwood, 2013). There are further modeling issues, such as whether to include random effects (REs)

at the teacher or school level (e.g., Gansle et al., 2012a). Once a model has been fit, the methodological decisions are not over. There remain a variety of methods that can be used to assess how much noise is present in the estimates, adjust for it, and address the issue of multiple tests.

In this article, we use an exceptionally large and diverse Texas dataset to estimate teacher quality differences between TPPs. We compare a variety of models, with clusters and random effects at various levels, and we compare a variety of methods for estimating the presence, size, and reliability of TPP differences.

We find that TPP point estimates are fairly robust to modeling decisions, but SE estimates are more sensitive and can be biased and volatile. While SE estimates are necessary for some purposes, we show that some methods can ignore the SE estimates and use the point estimates alone to estimate the variance that is due to true differences between TPPs and the variance that is due to noise. We also demonstrate graphical methods that can make the problems of noise and multiple tests more salient when TPP estimates are presented to policy makers.

In every plausible analysis, we find that the teacher quality differences between TPPs are small, and that estimates of those differences consist mostly of noise, even in large TPPs. We also find that few if any TPPs can be confidently flagged as different from average after adjustments are made for multiple tests. These results suggest that TPP accountability systems have very limited potential to improve student achievement (at least in the short run). In addition, careless use of TPP estimates can lead policy makers to make decisions about TPPs that are both arbitrary and ineffective.

2 Data

We use data from the Texas Education Agency (TEA) to estimate teacher quality differences between TPPs using student test scores in the spring of 2011. Although some Texas school districts had

previously linked teachers to students, 2011 was the first year for which TEA linked students to teachers statewide.

As the second largest U.S. state, Texas offers exceptional statistical power to detect even small TPP effects. The population of Texas exceeds the populations of Louisiana, New York City, Missouri, and Washington state combined. Table 1 shows that even a single year of Texas data, limited to 1st-3rd year teachers, has over 6,000 math teachers with nearly 300,000 students and nearly 5,000 reading teachers with over 200,000 students. If it is challenging to estimate TPP effects reliably in Texas, we may assume that it would be even more challenging in the 48 states that are smaller. A mid-sized state like Missouri, for example, would take five years to accumulate the sample size that we get from one year in Texas.

The data include math teachers from 95 TPPs and reading teachers from 92 TPPs. Texas TPPs are diverse in both size and approach. The largest TPP had 1,067 math teachers and 823 reading teachers in our data; the smallest TPP contributed only 3 reading teachers. Although many Texas TPPs are traditional programs run out of colleges and universities, the state's four largest TPPs are alternative TPPs, three of which are run for profit. Other alternative TPPs are run by independent school districts and regional educational service centers established by the state.

2.1 Test scores

Our dependent variables are high-stakes reading and math tests known as the Texas Assessment of Knowledge and Skills (TAKS). Texas students were required to take the TAKS in the springs of 2010, 2011, and before. The reading TAKS was given in 3rd-9th grades, and the math TAKS was given in 3rd-10th. TAKS was developed by Pearson Learning, which scaled scores using a one-parameter IRT model (DeMars, 2010). TAKS content was aligned with the state curriculum, and TAKS scores were more than

80% reliable and correlated positively with course grades (Texas Education Agency, 2011). We standardized TAKS scores within grade, and subject to facilitate interpretation.[1]

2.2 TPPs and other teacher variables

All student test scores were linked to the teacher who taught the tested subject in the year of the test. Students' math scores were linked to their math teacher, and their reading scores were linked to their reading teacher. In elementary school, a student's math and reading teacher were typically the same; in middle and high school, they were typically different.

Teachers were linked to the TPP that certified them in the tested subject. In our math model, teachers were linked to the TPP that certified them to teach math, and in our reading model, teachers were linked to the TPP that certified them to teach reading. 4th-8th grade teachers who were certified as "generalists" were treated as though they were certified to teach both math and reading. Teachers who were not certified in math, in reading, or as generalists were dropped from the analysis.

In addition to a teacher's TPP, our analysis included indicators for whether each teacher was in their first, second, or third year of teaching. This control is important because teachers improve with early experience (Papay & Kraft, 2015; Wiswall, 2013), and the distribution of teacher experience may be different for new and expanding TPPs than it is for older, established TPPs. Because TPP effects fade with time (Goldhaber et al., 2013), Texas law does not hold TPPs accountable for teachers after three years in the classroom (Texas State Legislature, 2009). We therefore excluded from our analysis teachers with more than three years' experience, as well as a few teachers who were certified before 2005 but started teaching more recently.

2.3 Student variables

Our models control for student-level covariates, including gender, race/ethnicity, limited English proficiency (LEP), and economic disadvantage (ED, which TEA defines as qualification for school meal

subsidies or other public assistance). We also coded variables summarizing the cumulative number of years that a student spent in ED or LEP status. Other student variables included indicators for special education status and the setting in which a special education student received instruction (mainstream or separate); indicators for whether the student had skipped or repeated a grade in the past 2 years; a measure of absenteeism, defined as the percentage of school days a student attended the school where they were tested; and two measures of mobility between schools: the number of schools in which the student was enrolled over the past four years, and the percentage of school days that the student was enrolled at their current school during the year of the test.

2.4 Classroom, school, and district variables

In addition to student variables, student test scores can be influenced by peer, classroom, school, and district characteristics that are beyond a teacher's control. To capture those influences, we coded a number of classroom, school, and district variables. At the classroom level, we calculated the class size and the percentage of students who were Hispanic, African American, ED, LEP, or in special education. We also calculated the average score of each classroom's students on the prior year's reading and math tests.

At the school level, we calculated the percentage of students who were ED, LEP, Hispanic, African American, or in special education, as well as the percentage of students who were referred for disciplinary problems in the previous year. We included indicators for whether the school was rural or suburban rather than urban, and an indicator for charter schools. To measure staff stability, we calculated the school's annual teacher turnover rate and the number of different principals who led the school over the past four years. Finally, we included the percentage of the schools' students who passed state reading and math tests, as well as indicators for how the school was rated in the state's

accountability system (exemplary, recognized, acceptable, unacceptable, with unrated as the omitted category). To avoid endogeneity, we lagged school pass rates and ratings by one year.

At the district level, we used the percentage of the district's budget that came from state rather than local funds. In Texas, as in many other states, state funding per pupil is higher in low-income districts (Corcoran, Evans, Godwin, Murray, & Schwab, 2004). We also included indicators for how the district was rated in the state's accountability system (exemplary, recognized, acceptable, unacceptable, with unrated as the omitted category). To avoid endogeneity, we lagged the district rating by one year.

3 Methods

3.1 Model

We fit a *lagged-score value-added model,* which regresses each student's test scores on their prior scores, an indicator for each TPP, and covariates. Lagged-score models are increasingly popular for estimating teacher value-added, and can easily be extended to estimate the average value-added of teachers from different TPPs. The econometric justification for a lagged-score model is that lagged scores proxy for the cumulative effects of prior school and non-school inputs, and therefore adjust for nonrandom assignment of students to teachers from different TPPs (Guarino, Reckase, & Wooldridge, 2014; Koedel, Mihaly, & Rockoff, 2015). Although the econometric assumptions of the lagged-score model are likely not perfectly met, simulations suggest that lagged-score models are more robust to nonrandom assignment than several other value-added models (Guarino et al., 2014). In addition, empirical results suggest that, at least in some data, lagged-score models can estimate teacher value-added with little bias (Chetty, Friedman, & Rockoff, 2014; Koedel, Mihaly, et al., 2015), although this claim has been challenged (Rothstein, 2014).

Our model for value added to reading scores is

$$
\begin{aligned}
Read_{yi} = \boldsymbol{\alpha}\,\boldsymbol{TPP}_t &+ \beta_1 Read_{y-1,i} + \beta_2 Math_{y-1,i} \\
&+ \beta_3 MaxRead_{y-1,i} + \beta_4 MaxMath_{y-1,i} \\
&+ \boldsymbol{\beta_5 Student}_i + \boldsymbol{\beta_6 Classroom}_c \\
&+ \boldsymbol{\beta_7 Teacher}_t + \boldsymbol{\beta_8 School}_s + \boldsymbol{\beta_9 District}_d \\
&+ e_i
\end{aligned}
\qquad (1),
$$

and our model for math scores is the same with $Math_{yi}$ as the dependent variable. The structure of the error term e_i can be modeled in several ways which we will discuss later.

The dependent variable $Read_{yi}$ (or $Math_{yi}$) represents the standardized score of individual student *i* on the reading (or math) test given in year *y*=2011. The lagged scores $Read_{y-1,i}$ and $Math_{y-1,i}$ are the same student's standardized scores on tests given in the prior year *y*–1=2010. We use lagged scores from two different subjects to reduce bias in estimating teacher value-added (Koedel, Mihaly, et al., 2015). Using longer lags—e.g., scores from years *y*–2 and *y*–3—may reduce bias as well (Koedel, Mihaly, et al., 2015), but it also introduces missing-data problems since many students lack scores at longer lags. In addition, since state testing begins in third grade, it is not possible to include lags of more than one year in the 4th-grade model, or lags of more than two years in the fifth-grade model. To adjust for ceiling effects, the model includes indicator variables $MaxRead_{y-1,i}$ and $MaxMath_{y-1,i}$ to flag the 3.5 percent of students who achieved the maximum possible score on the 2010 test. Other regressors include vectors of student, classroom, teacher, school, and district covariates, which we described in the Data section.

$\boldsymbol{TPP}_t$ is a column vector of indicators representing the *P* TPPs, and $\boldsymbol{\alpha} = [\alpha_1 \quad \alpha_2 \quad \cdots \quad \alpha_P]$ is a row vector representing the average value-added by teachers from each TPP. Because the model has an indicator for every TPP, it has no intercept, since an intercept would be collinear with the vector $\boldsymbol{TPP}_t$. In effect, each TPP has its own intercept. As a comparison to the *TPP model* in (1), we also fit a *no-TPP model* that had a single intercept and no TPP indicators.

We fit the models separately to each grade and to all grades together. In the all-grade model, we included grade indicators and let them interact with every regressor (except for the TPP indicators). These interactions allow for the possibility that the covariates had different coefficients in different grades. Similar all-grade TPP estimates can be obtained by averaging single-grade TPP estimates across grades.

3.2 Estimates and contrasts

From the TPP model, we get estimated TPP coefficients $\hat{\alpha}_p$, p=1,…,P, as well as *contrasts* $\Delta\hat{\alpha}_p$ which are defined as the difference between an individual TPP coefficient $\hat{\alpha}_p$ and the mean coefficient. The contrast vector $\Delta\hat{\alpha} = [\Delta\alpha_1 \ldots \Delta\alpha_P]$ has a covariance matrix V whose diagonal terms are squared SEs: $s_p^2 = SE^2(\Delta\hat{\alpha}_p)$.[2]

Some statistics center contrasts $\Delta\hat{\alpha}$ around the simple mean $\bar{\alpha}$ of the coefficients. Other statistics center student-weighted contrasts $\Delta\hat{\alpha}_n$ around a mean $\bar{\alpha}_n$ that weights each TPP coefficient $\hat{\alpha}_p$ by the number of students taught by teachers from that TPP. Other statistics use precision-weighted contrasts $\Delta\hat{\alpha}_{s2}$ centered around use a mean $\bar{\alpha}_{s2}$ that weights each coefficient by the inverse s_p^{-2} of its squared SE.

3.3 Clustered SEs

A vital issue is that the residuals e_i in equation (1) are correlated among students who are taught by the same teacher. If this correlation is ignored, then the SEs of the TPP estimates will be underestimated (Koedel, Parsons, et al., 2015).

One way to account for within-teacher correlation is to estimate teacher-clustered SEs (Koedel, Parsons, et al., 2015). Teacher-clustered SEs are estimated by calculating residuals around the OLS estimates, estimating the within-teacher covariance matrix of the residuals, and using that matrix to estimate the SE.

It may also seem plausible to cluster at a level higher than the teacher—such as the school, district, or TPP. In fact, it is common advice to cluster at the highest level possible (Cameron & Miller, 2015). Since clustered SEs can estimate arbitrary correlation structures, the idea is that clustered SEs at higher levels (e.g., schools, districts, or TPPs) will pick up not just correlations at higher levels but correlations at lower levels (e.g., teachers) as well.

There are some potential problems with using clustered SEs. One problem is that, if the residuals e_i are correlated, then OLS point estimates, though possibly unbiased, are not fully efficient. Another, more serious problem is that, if there are fewer than 40 clusters, clustered SEs are biased downward; that is, they tend to underestimate the true SEs (Cameron & Miller, 2015). In addition, with few clusters, clustered SEs are extremely *volatile* (a.k.a., variable, noisy, inefficient) in the sense that they fluctuate dramatically from one sample to another (Bell & McCaffrey, 2002).

In a TPP model, the bias and volatility of clustered SEs do not depend on the total number of clusters; instead they depend on the number of clusters *in each TPP*. Intuitively, each TPP's estimate depends primarily on clusters in that TPP, and the fact that one TPP has, say, 100 clusters does little to improve the SE for another TPP that has 5 clusters. This observation is not widely appreciated, but it is implicit in findings that the bias of clustered SEs is worse with skewed regressors—such as TPP dummies that are equal to one for only a small percentage of observations (Imbens & Kolesar, 2012; Pustejovsky & Tipton, 2016).

What this means is that, while teacher-clustered SEs may be reasonably accurate for large TPPs, teacher-clustered SEs will be biased and volatile for TPPs with fewer than 40 teachers. With fewer than 40 teachers, school- or district-clustered SEs will also be biased and volatile, since if a TPP has fewer than 40 teachers, those teachers will certainly be in fewer than 40 schools and fewer than 40 districts.

TPP-clustered SEs may be especially biased and volatile, since the SE of each TPP coefficient is estimated from a single cluster.

To address the bias and volatility of clustered SEs, a variety of methods have been developed, including bias reduced linearization and the wild cluster bootstrap (Bell & McCaffrey, 2002; Cameron, Gelbach, & Miller, 2008). We investigated these methods, but they did not solve our problem. First, as a practical matter, the current software implementations could not handle a dataset and model as large as ours. Second, even if the software could handle our data, it would not eliminate the problems of bias and volatility in SEs. The wild cluster bootstrap corrects significance levels but does not reduce bias or volatility in SEs (Cameron et al., 2008). Bias reduced linearization reduces bias but increases volatility (Bell & McCaffrey, 2002).

3.4 Teacher random effects

An alternative to clustered SEs is to model the correlated errors with teacher random effects (RE). A teacher RE model splits the residual into two components: $e_i = r_t + u_i$, where r_t is the teacher RE and u_i is the student residual. The RE model makes more assumptions than an OLS model with clustered SEs. While the clustered SE model makes no assumptions about the within-teacher covariance matrix, the teacher RE model assumes that, within teachers, e_i has a simple exchangeable correlation structure with an intraclass correlation of $\rho = \sigma_r^2/(\sigma_r^2 + \sigma_u^2)$, where σ_r^2 and σ_u^2 are the variances of r_t and u_i. Typically RE models also assume that r_t has a normal distribution, but RE estimates are often robust to non-normality (McCulloch & Neuhaus, 2011).

The choice between teacher REs and clustered SEs hinges on the RE assumptions and the size of the TPPs. If the RE assumptions are met, even approximately, then RE point estimates will be more efficient than OLS estimates, and RE SEs will be less biased (and less volatile) than clustered SEs, at

least in small TPPs (Green & Vavreck, 2008). On the other hand, if the RE assumptions are badly violated, then OLS estimates with clustered SEs may be preferable, at least for large TPPs.

3.5 School random effects vs. school fixed effects

In addition to teacher REs, we can add school REs and estimate a two-level hierarchical linear model (HLM). The inclusion of school REs may improve TPP coefficient estimates since much of the TPP variance lies between rather than within schools. A two-level HLM was used to estimate TPP coefficients in Louisiana (Gansle et al., 2012a).[3]

An alternative to school REs are school fixed effects (FEs). School FEs have the advantage of controlling for unobserved school-level covariates, but school FEs can be nearly collinear with some teacher TPPs (Mihaly et al., 2013). Because of the collinearity introduced by school FEs, some TPP estimates are identified only by a subset of new (i.e., 1st-3rd year) teachers who work in the same schools as new teachers from other TPPs. Although few cases are actually dropped from the regression, the number of teachers and schools that identify the coefficient of some TPPs can be very small. In the extreme, a TPP coefficient may not be estimable at all; more commonly, it will be estimable, but its SE will be very large. Bias can also occur if the teachers and schools that identify a particular TPP coefficient are not representative of the larger population.

3.6 Multiple comparisons

It is common practice to plot all the TPP contrasts $\Delta\hat{\alpha}_p$ with ordinary pointwise CIs (Boyd et al., 2009; Gansle et al., 2012a). And it is common to eyeball the CIs to see which ones do not cover zero, and interpret those TPPs as significantly different from the mean. This is equivalent to conducting P hypothesis tests.

The problem with this approach is that it fails to correct for multiple comparisons (Hsu, 1996). In Texas, for example, there are approximately P=100 different TPPs, and if we test each of them using a

.05 significance level (or a 95 percent CI), then we would expect to conclude that approximately five differ significantly from the average—even if all are in fact identical. The problem of multiple comparisons is exacerbated if we graph 68 percent CIs that extend only one SE in each direction (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2009; Gansle, Noell, & Burns, 2012a). If even 10 identical TPPs are compared using 68 percent CIs, there is a 98 percent chance ($1–.68^{10}$) of erroneously concluding that at least one TPP differs significantly from the average.

The simplest adjustment for multiple comparisons is the Bonferroni correction, which tests each hypothesis at a significance level of $.05/P$ or, equivalently, constructs CIs with a confidence level of $(1–.05/P)$x100 percent. This keeps the *familywise error rate* to five percent, meaning that, if all TPPs were identical, there would be approximately a five percent chance of erroneously concluding that at least one TPP differed from the average.

The Bonferroni correction is conservative, and less conservative corrections are available, including one that is tailored for our exact problem of making multiple comparisons with the mean (Fritsch & Hsu, 1997). But if the numbers of TPPs and teachers are large, as they are in Texas, the exact correction is practically indistinguishable from the Bonferroni correction, which is much easier to calculate. For example, with $P \geq 20$ TPPs and at least five teachers per TPP, the 95 percent Bonferroni intervals are only 0.3 percent wider than the exact intervals (Fritsch & Hsu, 1997). Our results use the Bonferroni correction; using the exact correction would not visibly change the results.

3.7 Definitions: Heterogeneity and reliability, homogeneity and the null distribution

The differences among the TPP point estimates are due partly to true *heterogeneity* between teachers from different TPPs, and partly to noise, or error in the estimates. The variance of the TPP estimates $\hat{\alpha}_p$ can be decomposed as follows

$$V(\hat{\alpha}_p) = \tau^2 + \sigma^2 \quad (2)$$

where $\tau^2 = V(\alpha_p)$ is the heterogeneity variance and $\sigma^2 = E\left(V(\hat{\alpha}_p - \alpha_p)\right)$ is the average variance of the estimation errors. The fraction of variance in $\hat{\alpha}_p$ that is due to heterogeneity rather than error is

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2} \quad (3).$$

We call ρ the *reliability* of the estimates $\hat{\alpha}$. Note that, for a given amount of estimation error, more heterogeneous estimates will also be more reliable.

If there is no heterogeneity, then the TPPs are *homogeneous* and their estimates are completely unreliable; they differ from one another only because of estimation error. The null hypothesis of homogeneity can be defined in several equivalent ways:

$$\begin{aligned} H_0&: \alpha_1 = \alpha_2 = \cdots = \alpha_P \\ \text{or } H_0&: \tau^2 = 0 \\ \text{or } H_0&: \rho = 0 \end{aligned} \quad (4)$$

Under H_0, the estimates $\Delta\hat{\alpha}_p$ would still vary because of estimation error. The distribution of estimates under H_0 is the *null distribution* $\mathcal{D}_0$, and we can describe $\mathcal{D}_0$ as follows. Under H_0, each $\Delta\hat{\alpha}_p$ would have an asymptotic normal distribution with a mean of zero and a variance estimated by $\hat{s}_p^2$, p=1,…,P. It follows that $\mathcal{D}_0$ is an equal mixture of P independent[4] normal distributions with means of zero and different variances. We approximate the null distribution using the following procedure. For the p^{th} TPP, the null distribution is $N(0, \hat{s}_p^2)$, from which we draw the 1st through 99th percentiles $\{q_{1,p}, \ldots, q_{99,p}\}$. Then for all the TPPs together, we approximate the null distribution $\mathcal{D}_0$ with a set containing all the percentiles that we have drawn for the individual TPPs—i.e., $\widehat{\mathcal{D}}_0 \approx \{q_{1,1}, \ldots, q_{99,1}, q_{1,2}, \ldots, q_{99,2}, \ldots q_{1,P}, \ldots, q_{99,P}\}$.[5]

Under H_0, the TPP contrasts $\Delta\hat{\alpha}_p$ would approximate the (P+1)-quantiles from $\widehat{\mathcal{D}}_0$ (e.g., the deciles if P=9, or the percentiles if P=99). We call these the *null quantiles*, or *noise quantiles*. By plotting the

noise quantiles over the observed $\Delta\hat{\alpha}_p$ values, we can visually compare the observed distribution to the noise distribution. If the observed distribution and the noise distribution are similar, we can conclude that most of the variation in the estimates is due to noise. If the observed distribution is more dispersed than the noise distribution, we have visual evidence that there are real teacher quality differences between TPPs.

3.8 Tests and estimates of heterogeneity and reliability

How can we test the null hypothesis of homogeneity and estimate the heterogeneity variance τ^2 and the reliability ρ? We review the statistics that have been used previously and point out that they are sensitive to biases in the SE estimates. We then propose new statistics that do not require SE estimates at all.

3.8.1 Homogeneity tests

One way to test the null hypothesis of homogeneity is with a likelihood ratio statistic *LR* that compares the log-likelihoods ℓ of the TPP and no-TPP models:

$$LR = 2(\ell_{TPP} - \ell_{noTPP}) \qquad (5)$$

Under the null hypothesis of homogeneity, *LR* follows a χ^2_{P-1} distribution if the sample is large and the model is correctly specified. The LR test is asymptotically most powerful, but it can only be used with maximum likelihood (ML) estimates, and even then it cannot be used with clustered SEs, because the likelihood ignores the clustering. In addition, if we want estimates and SEs from a subset of TPPs, such as the large ones, we can only calculate *LR* if we re-estimate the model on that subset.

A simpler test statistic is Cochran's *Q*, which compares the squared contrasts to their squared SEs (Cochran, 1954; Koedel, 2009; Koedel, Parsons, et al., 2015):

$$Q = \sum_{p=1}^{P} \frac{\Delta\hat{\alpha}_{s2,p}^2}{\hat{s}_p^2} \quad (6)$$

Q has the same null distribution as *ML*, but *Q* can be used with any type of estimate (ML, GLS, etc.) with or without clustered errors. *Q* can also be calculated from any set of estimates and SEs from a single model run. There is no need to compare the TPP and no-TPP models, and no need to rerun the model on a subset of TPPs.

A related test is the Wald statistic $W = \Delta\hat{\alpha}_n^T \hat{V}^{-1} \Delta\hat{\alpha}_n$, which is like *Q* except that *W* uses the whole covariance matrix $\hat{V}$ while *Q* only uses the diagonal elements (the squared SEs) (Koedel, 2009; Koedel, Parsons, et al., 2015). In our TPP model, *W* and *Q* are typically very similar because the off-diagonal elements of $\hat{V}$ are close to zero.[6] In some situations, though, we found that *W* was less robust than *Q*,[7] so we omit *W* from our results.

3.8.2 Reliability and heterogeneity estimates

Reliability can be estimated by

$$\hat{\rho}_Q = 1 - \frac{P-1}{Q} \quad (7)$$

$\hat{\rho}_Q$ is called I^2 in meta-analysis, where it is often reported with a test-based CI (Higgins & Thompson, 2002). $\hat{\rho}_Q$ can be multiplied by the variance of the point estimates $\hat{V}(\hat{\alpha}_p)$ to get an estimate of the heterogeneity variance (Koedel, 2009):

$$\hat{\tau}_Q^2 = \hat{\rho}_Q \hat{V}(\hat{\alpha}_p), \text{where } \hat{V}(\hat{\alpha}_p) = \frac{1}{P-1}\sum(\hat{\alpha}_p - \overline{\alpha})^2 \quad (8)$$

Another estimate of the heterogeneity variance is the difference between the variance of the TPP point estimates and the variance of the null distribution.

$$\hat{\tau}_H^2 = \hat{V}(\hat{\alpha}_p) - \hat{V}(\mathcal{D}_0), \text{where } \hat{V}(\mathcal{D}_0) = \frac{1}{P}\sum \hat{s}_p^2 \quad (9)$$

(Aaronson, Barrow, & Sander, 2007; Cochran, 1954; Hedges, 1983). $\hat{\tau}_H^2$ can be divided by the variance of the point estimates to get another estimate of reliability:

$$\hat{\rho}_H = \hat{\tau}_H^2 / \hat{V}(\hat{\alpha}_p) \quad (10)$$

The estimate $\hat{\tau}_H^2$ is unbiased, but $\hat{\rho}_Q$ has a slight negative bias if the number of TPPs is small (von Hippel, 2015). Any heterogeneity or reliability estimate can be negative, and it is customary to round negative estimates up to zero. Rounding up yields a positive bias if the true heterogeneity is close to 0 and the number of TPPs is small (von Hippel, 2015).

Another heterogeneity estimator is $\hat{\tau}_{EB}^2$, which is the variance of the empirical Bayes (EB) contrasts (e.g., Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2009; Goldhaber et al., 2013; Koedel, Parsons, et al., 2015). The EB contrasts are obtained by multiplying each ordinary contrast $\Delta\hat{\alpha}_p$ by an estimate of its reliability (Merrmann, Walsh, Isenberg, & Resch, 2013).

Unfortunately, $\hat{\tau}_{EB}^2$ has a substantial bias even if the number of TPPs is large. To see the bias simply, suppose that every TPP contrast $\Delta\hat{\alpha}_p$ has known reliability ρ.[8] Then the EB contrasts are $\rho\Delta\hat{\alpha}_p$ with variance $\hat{\tau}_{EB}^2 = \rho^2\hat{V}(\Delta\hat{\alpha}_p)$. However, the true heterogeneity variance is $\tau^2 = \rho V(\Delta\hat{\alpha}_p)$. So $\hat{\tau}_{EB}^2$ underestimates τ^2 by a factor of ρ. We do not recommend $\hat{\tau}_{EB}^2$ and will not present it in our Results.

3.8.3 Using TPP point estimates

The statistics above rely on SE estimates, but SEs are tricky to estimate, and some popular SE estimators are biased, as we will see.

We now propose alternative statistics that do not require accurate SE estimates. These statistics work by comparing different point estimates of the TPP coefficients. From our models we get TPP point estimates for 2 different subjects (reading, math) and 6-7 different grades (4^{th}-9^{th} in reading, 4^{th}-10^{th} in math). In different data, we could also get TPP point estimates for different school years or for different cohorts of teachers.

If we have two sets of independent and exchangeable TPP estimates, then the correlation between them estimates the reliability ρ, and the covariance estimates the heterogeneity variance τ^2. If the correlation (and covariance) are significantly greater than zero, then we can reject the null hypothesis of homogeneity.

If we have more than two sets of TPP estimates, then the bivariate correlation generalizes to the intraclass correlation, which can be estimated using analysis of variance (ANOVA). With *J* independent estimates for each TPP, the ANOVA model is

$$\Delta\hat{\alpha}_{pj} = \Delta\alpha_p + u_{pj} \qquad (11),$$

where $\Delta\hat{\alpha}_{pj}$ is the j^{th} estimated contrast for TPP p in grade g, $\Delta\alpha_p$ is the true contrast, and u_{pj} is random estimation error. The null hypothesis of homogeneity is tested by the ANOVA *F* statistic. Standard ANOVA formulas[9] (Fisher, 1925) give the between-group variance, which we call $\hat{\tau}^2_{ICC}$ and interpret as an estimate of the heterogeneity variance. Standard formulas also give the intraclass correlation *r*, which estimates the reliability of a single TPP estimate. If *J* TPP estimates are averaged together—for example, if we average TPP estimates across *J* grades—then the reliability of the average is estimated by (Winer, Brown, & Michels, 1991)

$$\hat{\rho}_{\text{ICC}} = \frac{Jr}{1 + (J-1)r} \qquad (12)$$

These formulas assume that the TPP estimates are independent and exchangeable. If estimates are not independent, then the formulas will overestimate both heterogeneity and reliability, and we will reject the null hypothesis of homogeneity more often than we should. If estimates are independent but not exchangeable, then we will underestimate reliability, and we will reject homogeneity less often than we should.

The assumptions of independence and exchangeability are most plausible when we are comparing the same subject taught by teachers in similar grades. For example, 4th and 5th grade math teachers are

independent and exchangeable, while 4th and 10th grade math teachers are independent but may not be exchangeable if different skills are needed to teach 4th vs. 10th grade math.

Assumptions can be violated in other situations. For example, independence is violated if we correlate TPP estimates across two different school years, but many of the teachers are the same in both years. If we correlate reading and math estimates, then the estimates may be independent but they are not exchangeable if TPPs are more heterogeneous in math than in reading, or if some TPPs' reading teachers are better or worse than their math teachers.

The assumptions of independence and exchangeability may seem strict, but they are not limited to the ANOVA approach described above. *Any* heterogeneity estimator assumes independence and exchangeability when it combines data across different grades, subjects, or cohorts.

4 Results

4.1 Illustrative TPP estimates

Figure 1 displays caterpillar plots of TPP contrasts from our all-grade teacher RE models, with 95 percent pointwise CIs and a reference line at the average of zero. We illustrate the noisiness of the estimates by overlaying the null distribution, which shows what the distribution of TPP contrasts would look like if there were no true differences between TPPs and nothing but estimation error were present. The null distribution is barely less dispersed than the observed distribution, suggesting that the observed distribution of TPP contrasts consists primarily of noise, with little signal.

Both the observed distribution and the null distribution have a sideways S shape. In the value-added community, a sideways S is sometimes interpreted as meaning that most TPPs are very similar, while a few, in the tails, are very bad or very good. But clearly that interpretation is wrong since the null distribution, which assumes no TPP differences, also has an S shape. The reason for the null

distribution's S shape is that estimation error has a normal mixture distribution, and the cumulative distribution function for a normal mixture is S-shaped.

To decide whether a TPP is significantly better or worse than average, it is common to plot a 95 percent pointwise CI around each estimate. Figure 1 does this, but the practice is misleading. It is tempting to infer that a TPP is different from average if its pointwise CI does not cross zero, but this is not necessarily the case. Even if there were no true differences between TPPs, 5 percent of 95 percent pointwise CIs—or about 9-10 of the 187 intervals in Figure 1—would not cross the reference line. This is the problem of multiple comparisons.

To correct for multiple comparisons, Figure 1 includes 95 percent Bonferroni CIs that adjust for the fact that we have 95 TPPs in math and 92 in reading. Looking for Bonferroni intervals that do not cross the reference line, we conclude that no TPPs are significantly different from average in math, and only one TPP is significantly different (better) than average in reading.[10]

4.2 Model sensitivity

The estimates in Figure 1 came from a teacher RE model, and some of our conclusions would change if we fit a different model. Table 2a summarizes the distribution of TPP point estimates and SE estimates under different models.

The model with school FEs (and teacher REs) appears to produce the worst estimates. Its SEs are about 50% larger on average, and its point estimates are about 50% more variable (according to the SD) than the point estimates and SEs obtained from other models. School FE estimates have very low correlations (.14-.16) with other estimates in reading, and only moderate correlations with other estimates in math (.50-.60). There is one TPP for which the school FE model fails to provide an estimate at all. These are all symptoms of the fact that school FE estimates are identified by a limited and possibly unrepresentative sample of schools and teachers (Mihaly et al., 2013).

OLS point estimates are more efficient, but it is tricky to estimate SEs under OLS. OLS SE estimates are too small if they are not clustered (Koedel, Parsons, et al., 2015), and SE estimates are even smaller if they are clustered at the TPP level, according to Table 2a. TPP-clustered SEs are biased downward because they have only one cluster for each TPP, and about 40 clusters per TPP are needed to avoid downward bias (Cameron & Miller, 2015). SEs are much larger if they are clustered at the teacher, school, or district level, but even these SEs have some downward bias for small TPPs with fewer than 40 teachers, schools, or districts. Notice that the district-clustered SEs are slightly smaller than the school- or teacher-clustered SEs. This is because there are fewer districts than schools or teachers.

SE estimates from a teacher RE model are slightly larger, and probably less biased, than SE estimates from an OLS model with teacher clustering (or school or district clustering). Figure 2 shows how TPP size affects SE estimates with teacher clustering vs. teacher REs. With teacher REs, SE estimates decrease smoothly as the inverse square root of the number of teachers. Teacher-clustered SEs are similar when the number of teachers is greater than 40. But with fewer than 40 teachers, teacher-clustered SEs are too small on average and extremely volatile, with no smooth relationship to TPP size. This reflects the volatility and downward bias of clustered SEs when there are few clusters (Bell & McCaffrey, 2002).

While teacher REs improve the SE estimates, they do not change the point estimates very much. Teacher RE point estimates are highly correlated with OLS estimates (.89 in math and.97 in reading), and just slightly more dispersed. The estimates also change little if we add school REs to the teacher REs. TPP estimates from a model with school and teacher REs are highly correlated (.97-.98) with teacher RE point estimates, and slightly less dispersed.

The last two columns of Table 2a show how many TPPs are estimated to be significantly different from the mean (at $p<.05$). Two patterns are evident. First, if a model underestimates the SEs, it will

suggest that a spuriously large number of TPPs differ significantly from the average. The number of significant differences is substantially exaggerated under OLS with no clustering and under OLS with TPP clustering. The number of significant differences is also slightly exaggerated under OLS with district clustering.

The second pattern is that even if a model produces reasonable SE estimates, we will still overstate the number of significant differences if we fail to correct for multiple tests. Consider the model with teacher REs. Without correction this model suggests that 10 TPPs differ significantly from the average in math, and 11 TPPs differ significantly from the average in reading. But after the Bonferroni correction for multiple tests, only 0 TPPs differ significantly from the average in math, and only 1 differs significantly from the average in reading.

4.3 Large TPPs

Table 3b summarizes point estimates, SEs, and significant differences for TPPs with at least 40 reading or math teachers in our data. These larger TPPs represent only 40-50 percent of the TPPs in Texas, but they train 80 percent of the state's new teachers. Both policy and statistical arguments can be made for limiting accountability to large TPPs. From a policy point of view, the decision to shut down or expand a large TPP will affect a larger number of students. Statistically, large TPPs have more precise point estimates and less danger of bias in SE estimates; in addition, there are fewer large TPPs, and this reduces the danger of multiple comparisons.

Some estimation methods do not improve in large TPPs. School FE estimates are still very imprecise and may be biased as well. OLS and TPP-clustered SEs still have a severe downward bias.

Other estimation methods do improve in large TPPs. In large TPPs, teacher-, school-, and district-clustered SEs have little bias, and are very similar to SEs from teacher RE models (with or without school REs) The correlation between OLS and teacher RE point estimates remains strong in large TPPs

(.85 in math, .84 in reading), though not as strong as it was when all TPPs were included. The correlation between teacher RE and teacher+school RE models is strong as well (.96 in math, .94 in reading).

Despite the smaller SEs of large TPPs, and despite the fact that a sample limited to large TPPs has fewer comparisons to correct for, it remains difficult (but not impossible) to single out specific TPPs as significantly different from average. If we apply the Bonferroni correction and limit our attention to the teacher RE estimates, only one TPP differs significantly from the average in reading, and only one differs significantly from the average in math. It is the same TPP that stands out in both reading and math, suggesting that the Bonferroni correction may have helped us to find a TPP that is truly different from average. The TPP has an estimated contrast of .07 in both reading and math, though after EB shrinkage the contrast shrinks to just .04 in math and .01 in reading.

4.4 Heterogeneity and reliability

In light of the previous section's difficulty highlighting individual TPPs that are significantly different, some readers might doubt that there are any teacher quality differences between TPPs at all. This section suggests that there are differences, but they are very small and not very reliably estimated.

Table 3a tests and estimates the heterogeneity and reliability of the contrasts for all TPPs. These heterogeneity and reliability estimates are calculated from the SE estimates, so if the SE estimates are biased, the heterogeneity and reliability estimates will be biased as well. In particular, if SE estimates are too low—as they are under unclustered OLS or under TPP clustering—then estimates of reliability and heterogeneity will be too high, and hypothesis tests will reject homogeneity more often than they should. For that reason, we give no weight to the OLS or TPP-clustered estimates and relatively little weight to the teacher-, school-, or district-clustered estimates unless the TPPs are large. We also give

little weight to the school FE estimates, which are very imprecise and possibly biased. Instead, we favor the teacher RE estimates, with or without school REs.

The teacher RE estimates in Table 3a suggest that we can reject the null hypothesis of homogeneity; the *p* values for the *Q* and *LR* tests are less than.05. The TPP reading estimates have a heterogeneity SD of .02-.03 and are most likely 18 to 33 percent reliable. The TPP math estimates have a heterogeneity SD of .04-.05 and are most likely 23 to 40 percent reliable. These estimates are highly uncertain, however; the confidence intervals for reliability range from 0 to 52 percent.

We initially hoped that large TPPs would have more reliable estimates, but they do not. According to Table 3b, the large-TPP contrasts are just 18 to 46 percent reliable in math and 0 to 36 percent reliable in reading. These low reliabilities occur because, although large TPP contrasts are more precise, they also appear to be less heterogeneous. The large TPPs have an estimated heterogeneity SD of just 0.01-.02 in math and 0-.01 SD in reading. Under the model with teacher REs and school REs, we cannot even reject the null hypothesis of homogeneity ($p>.05$ in both math and reading).

Again, this summary is based on models with teacher RE and possibly school REs. The models with clustered SEs at the school, teacher, or district level tell a similar but not identical story, at least in large TPPs.

Given the sensitivity of SE estimates, it may be helpful to ignore the SEs and estimate reliability and heterogeneity by correlating point estimates across grades (see section 3.8.3 in the Methods). The point estimates from different models are similar, even when the SE estimates are different. For example, OLS point estimates stay the same regardless of how SEs are clustered. And OLS point estimates are strongly correlated with teacher RE point estimates, though school FE point estimates are different.

Table 4a uses the point estimate approach to estimate reliability and heterogeneity. The estimates for heterogeneity are low. Across all TPPs, the estimated heterogeneity SD is 0.03 in math and 0.02 in reading. Among large TPPs, the estimated heterogeneity SD is .01-.02 in math and .01 in reading. Some reliability estimates are low as well; among large TPPs in reading, the reliability estimate is just 10 percent, and we cannot reject the null hypothesis that there is no reliability at all. The estimates are quite consistent across models, with the exception of the school FE models, which we discount because their estimates are imprecise and possibly biased.

5 Conclusion

5.1 How large are TPP differences? How reliable? Which TPPs are different?

In the introduction we argued that, for TPP accountability to increase student performance, several conditions must be met.

1. The differences between TPPs must be consequentially large.
2. It must be possible to estimate those differences reliably.
3. It must be possible to single out individual TPPs that are better or worse than average.

We can now assess those conditions by answering the three questions in the title

Question 1. How large are the differences between TPPs? While most of our results suggest that real differences between TPPs exist, the differences are not large. Our estimates vary a bit with our statistical methods, but averaging across plausible methods we conclude that between TPPs the heterogeneity SD is about .03 in math and .02 in reading. That is, a 1 SD increase in TPP quality predicts just a .03 SD increase in student math scores and a .02 SD increase in student reading scores.

The differences between TPPs are not large in an absolute sense, and also not large when compared to other differences between groups of teachers. For comparison, using the same value-added

model, we estimate that the average difference between 1st and 2nd year teachers is 0.04 SD in student math scores and 0.03 SD in student reading scores. So a 2nd year teacher from an average TPP is probably better than a 1st year teacher from a good TPP.

Question 2. How reliable are TPP estimates? Even if the differences between TPPs were large enough to be of policy interest, accountability could only work if TPP differences could be estimated reliably. And our results raise doubts that they can. Every plausible analysis that we conducted suggested that TPP estimates consist mostly of noise. In some analyses, TPP estimates appeared to be about 50 percent noise; in other analyses, they appeared to be as much as 80 or 90 percent noise, despite our Texas-sized sample. Even in large TPPs the estimates were mostly noise, because the differences between large TPPs, though more precisely estimated, were also smaller than the differences between small TPPs.

It is plausible (though it needs to be assessed empirically) that TPP estimates would be more reliable if we had more than one year of data. But if several years of data are required to obtain reliable TPP estimates in Texas, what does that imply for other states? A mid-sized state like Missouri would require 5 years to accumulate the amount of data that we get from a single year in Texas.

TPP estimates are sensitive and uncertain. The estimates are noisy even if we settle on a single model, and there is also uncertainty about which model to fit. While all TPP evaluations to date have used a lagged-score value-added model, evaluators have made different decisions about clustering, REs, and FEs, and we have shown that some of those decisions have major consequences for the resulting estimates, especially in small TPPs. In addition, different evaluations have used different sets of covariates, and the choice of covariates can change the distribution of TPP estimates as well (Lincove et al., 2014).

It is possible that some TPP estimates are not just uncertain but also biased. Possible sources of bias include model misspecification and nonrandom assignment of TPPs' teachers to schools and students. The biases of lagged-score value-added models are small when compared to differences between teachers (Chetty et al., 2014; Koedel, Mihaly, et al., 2015), but the biases may be larger when compared to the smaller differences between TPPs.

Question 3. Which TPPs are different? Even if we focus on estimates from a single model, it remains hard to identify which TPPs differ from the average. It is not just that TPP differences are small and our estimates of them are uncertain—there is also the problem of multiple comparisons. Before we correct for multiple comparisons, many TPPs appear significantly different, but after we correct for multiple comparisons, just 0-2 TPPs appear significantly different from the average. If we restrict accountability to large TPPs, we have fewer comparisons to make, but it is no easier to detect significant differences because the differences between large TPPs, at least in Texas, are very small.

We can radically reduce the number of comparisons if we combine TPPs and ask broader questions, such as whether alternative TPPs produce better teachers than traditional TPPs (Kane, Rockoff, & Staiger, 2008), whether for-profit TPPs produce better teachers than nonprofit TPPs (Lincove, Osborne, Mills, & Bellows, 2015), or whether TPPs that involve students in teaching practice produce better teachers than TPPs that don't (Boyd et al., 2009). These are interesting questions, but from a policy point of view, they are fundamentally different than the accountability problem of identifying which individual TPPs are better or worse. For example, even if teachers from alternative TPPs were on average better than those from traditional TPPs, we could not justify shutting down all traditional TPPs. There might be some traditional TPPs that are excellent.

5.2 How general are our results?

Our finding that there are only small teacher quality differences between TPPs may seem surprising at first. After all, TPPs differ substantially both in selectivity and in their approach to teacher training. Some TPPs accept only 10 percent of applicants, while others take nearly all comers. Some TPPs are 4-year degree programs, while others last as little as 12 weeks. Yet we find only small teacher quality differences between TPPs in Texas, and similar results have been obtained in Missouri and Washington state (Goldhaber et al., 2013; Koedel, Parsons, et al., 2015). It is a little surprising that differences in TPP selectivity and training don't produce bigger differences in teacher effectiveness.

Yet results like this are common in education research. In many areas of education, little of the variation in individual success lies between institutions. In elementary school, only 20 percent of the variation in student test scores lies between schools (Coleman et al., 1966). Among college graduates with the same major, only 1 to 9 percent of the variance in log earnings lies between graduates of different colleges (Rumberger & Thomas, 1993). Among PhD economists, only 10 percent of the variance in research productivity, lies between graduates of different PhD programs (Conley & Önder, 2014).[11] Perhaps we should not be surprised by results suggesting that only 1 to 3 percent of the variance in teacher quality lies between teachers from different TPPs (Goldhaber et al., 2013; Koedel, Parsons, et al., 2015). And since the total heterogeneity among teachers is .09 to .16 SD in student test scores (Staiger & Rockoff, 2010), it stands to reason that the heterogeneity between TPPs would be as small as .01 to .03 SD.[12]

It is possible that results would be different for different outcome variables. Most TPP evaluations have focused exclusively on reading and math scores, although one evaluation also looked at science and social studies scores (Gansle, Noell, & Burns, 2012). It would be informative to estimate between-TPP differences in teacher attrition and in teacher effects on grade retention and graduation rates. In fact,

policy often highlights graduation as an outcome that TPPs should be accountable for (Levine, 2006; Texas State Legislature, 2009; US Department of Education, 2011).

5.3 Recommended methods

Although our results suggest limits on the potential of TPP accountability systems, implementation of these systems may continue as the merits of the policy are debated. For evaluators who continue to estimate teacher quality differences between TPPs, we have some recommendations and cautions regarding which methods to use.

While TPP point estimates are somewhat similar across models, SE estimates are more sensitive and can be biased and volatile. If there are several sets of independent point estimates—e.g., estimates from different grades, or estimates in different subjects—then we can ignore the SE estimates and estimate heterogeneity and reliability using the point estimates alone. However, we need SE estimates to evaluate which TPPs are significantly different from average.

To estimate SEs, it is essential to account for the correlation between students taught by the same teacher. Within-teacher correlation can be modeled using either teacher clusters or teacher REs. Teacher clusters and teacher REs give similar SE estimates for large TPPs with at least 40 teachers. For smaller TPPs, though, teacher REs are preferable because teacher teacher-clustered SEs are volatile and biased. The bias of clustered SEs does not improve if we cluster at the school or district level instead of the teacher level, and if we cluster at the TPP level the bias gets much worse.

When using an RE model, there is a statistical case for adding REs at the school and district level as well as the teacher level. These higher-level REs make only a small difference to the TPP estimates, but the difference can be large enough to nudge some TPP estimates from significance to insignificance.

TPP estimates are typically compared using a caterpillar plot, but we argue that traditional caterpillar plots are misleading in two ways. First, caterpillar plots rank TPPs by their estimated effects,

and it is easy to get the impression that the TPPs are being ranked on quality, even though the estimates consist primarily of noise. Traditional caterpillar plots can also mislead users by using ordinary pointwise CIs, which are too narrow because they ignore the problem of multiple comparisons.

To highlight the issues of noise and multiple comparisons, caterpillar plots should use Bonferroni CIs and overlay a null distribution that shows what the estimates would look like if only estimation error were present and there were no real differences between TPPs. Highlighting noise and correcting for multiple comparisons may help to steer policymakers away from unnecessary or counterproductive actions such as such as closing an average TPP because a noisy estimate makes it appear worse than it is. In addition, cautious analysis can highlight the occasional situation where—despite noise and multiple comparisons—we can have confidence that one TPP is better or worse than average.

Endnotes

[1] We excluded students who took the "accommodated" TAKS (for special education students) or the Spanish-language TAKS in 2010 or 2011. The fraction of students who took the Spanish TAKS is surprisingly small. By 4th grade, only 6 percent of Texas students took the Spanish reading test, and 3 percent took the Spanish math test. By 5th grade, only 3 percent took the Spanish reading test, and 1 percent took the Spanish math test.

[2] We obtained both $\Delta\hat{\alpha}$ and $\hat{V}$ using Stata's postestimation command *contrast gw.TPP*. An alternative to calculating contrasts would be to mean-center all the regressors, including the dummies.

[3] The article by Gansle et al. (2012) describes a model with random effects at the school level and the "teacher/classroom" level. The term "teacher/classroom" is ambiguous since a teacher can have more than one classroom. On July 22, 2015, we wrote to Gansle about this ambiguity and she replied that, "The random effect was really a teacher effect across one to several classes."

[4] Here we are assuming that the correlations among the estimates are small. As remarked earlier, this assumption is reasonable when *W* is similar to *Q*.

[5] We implemented this approximation procedure in a few lines of Stata code, and compared the results to quantiles from the exact distribution $\mathcal{D}_0$ which we calculated using Mathematica software. The results were visually indistinguishable.

[6] In an OLS model where the only regressors were TPP dummies, the covariance matrix of the TPP estimates would be $V = \sigma^2(TPP^TTPP)^{-1}$, whose off-diagonal elements are zero. Covariates, random effects, and clustering may add a little to the off-diagonal elements.

[7] For most models, *W* was close to *Q* (within 0-7%) but with district clustering *W* was approximately twice as large as *Q*, and with TPP clustering *W* was 30,000-40,000 times larger than *Q* (e.g., $W \approx 10^8$ when Q≈3,000). The latter result seemed implausible, although it doesn't matter since we will show there are good reasons not to use TPP clustering.

[8] In practice, each contrast has a different reliability ρ_p, which is not known but is estimated by $\hat{\rho}_p = \hat{\tau}^2/(\hat{\tau}^2 + \hat{s}_p^2)$, where $\hat{\tau}^2$ is an estimate of τ^2 (Merrmann, Walsh, Isenberg, & Resch, 2013). What this means is that we need an estimate of τ^2 before we can calculate the EB estimates. There is something circular about then using the variance of the EB estimates as a new estimate of τ^2.

[9] ANOVA calculations are implemented by the *loneway* command in Stata.

[10] Notice that the single significant reading estimate does not have the largest absolute point estimate in the caterpillar plot. It only has the largest point estimate relative to its SE.
[11] We calculated this fraction of variance by running an ANOVA on the data published by Conley and Önder (2014). Conley and Önder summarize their results in a different way.
[12] To walk through the calculation: if the SD between teachers is .09 and only 1 percent of the teacher variance (SD^2) lies between TPPs, then the SD between TPPs would be $.09 \times \sqrt{.01} \approx .01$. Alternatively, if the SD between teachers is .16 and as much as 3 percent of the teacher variance lies between TPPs, then the SD between TPPs would be $.16 \times \sqrt{.03} \approx .03$.

Published in 2016 by *Economics of Education Review* 53:31-45.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics*, *25*(1), 95–135. http://doi.org/10.1086/508733

Bell, R. M., & McCaffrey, D. F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, *28*(2), 169–182.

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.

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. *The Review of Economics and Statistics*, *90*(3), 414–427.

Cameron, A. C., & Miller, D. L. (2015). A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, *50*(2), forthcoming.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates. *The American Economic Review*, *104*(9), 2593–2632. http://doi.org/10.1257/aer.104.9.2593

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*, 101–129.

Coleman, J. S., Campbell, E. Q., Hobson, C. F., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of Educational Opportunity*. Washington, DC: Department of Health, Education and Welfare.

Conley, J. P., & Önder, A. S. (2014). The Research Productivity of New PhDs in Economics: The Surprisingly High Non-Success of the Successful †. *Journal of Economic Perspectives*, *28*(3), 205–216. http://doi.org/10.1257/jep.28.3.205

Corcoran, S., Evans, W. N., Godwin, J., Murray, S. E., & Schwab, R. M. (2004). The changing distribution of education finance, 1972-1997. In K. M. Neckerman, (Ed.), *Social Inequality*. New York, NY: Russell Sage Foundation.

DeMars, C. (2010). *Item Response Theory*. Oxford University Press, USA.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. London: Oliver and Boyd.

Fritsch, K. S., & Hsu, J. C. (1997). Multiple Comparisons With the Mean. In S. Panchapakesan & N. Balakrishnan (Eds.), *Advances in Statistical Decision Theory and Applications* (pp. 189–204). Birkhäuser Boston.

Gansle, K. A., Noell, G. H., & Burns, J. M. (2012). Do Student Achievement Outcomes Differ Across Teacher Preparation Programs? An Analysis of Teacher Education in Louisiana. *Journal of Teacher Education*, *63*(5), 304–317. http://doi.org/10.1177/0022487112439894

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. http://doi.org/10.1016/j.econedurev.2013.01.011

Green, D. P., & Vavreck, L. (2008). Analysis of Cluster-Randomized Experiments: A Comparison of Alternative Estimation Approaches. *Political Analysis*, *16*(2), 138–152. http://doi.org/10.1093/pan/mpm025

Guarino, C. M., Reckase, M. D., & Wooldridge, J. M. (2014). Can Value-Added Measures of Teacher Performance Be Trusted? *Education Finance and Policy*, *10*(1), 117–156. http://doi.org/10.1162/EDFP_a_00153

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin*, *93*(2), 388–395. http://doi.org/10.1037/0033-2909.93.2.388

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. http://doi.org/10.1002/sim.1186

Imbens, G. W., & Kolesar, M. (2012). *Robust Standard Errors in Small Samples: Some Practical Advice* (Working Paper No. 18478). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w18478

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*, *27*(6), 615–631. http://doi.org/10.1016/j.econedurev.2007.05.005

Koedel, C. (2009). An empirical analysis of teacher spillover effects in secondary school. *Economics of Education Review*, *28*(6), 682–692. http://doi.org/10.1016/j.econedurev.2009.02.003

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*. http://doi.org/10.1016/j.econedurev.2015.01.006

Koedel, C., Parsons, E., Podgursky, M., & Ehlert, M. (2015). Teacher Preparation Programs and Teacher Quality: Are There Real Differences Across Programs? *Education Finance and Policy*, *10*(4), 508–534. http://doi.org/10.1162/EDFP_a_00172

Levine, A. (2006). *Educating School Teachers* (No. 2). Washington, DC: The Education Schools Project. Retrieved from http://www.edschools.org/teacher_report_release.htm

Lincove, J. A., Osborne, C., Dillon, A., & Mills, N. (2014). The Politics and Statistics of Value-Added Modeling for Accountability of Teacher Preparation Programs. *Journal of Teacher Education*, *65*(1), 24–38. http://doi.org/10.1177/0022487113504108

Lincove, J. A., Osborne, C., Mills, N., & Bellows, L. (2015). Training Teachers for Profit or Prestige: The Effects of Market and Institutional Incentives of Teacher Preparation Programs on Student Performance. *Journal of Teacher Education*.

McCulloch, C. E., & Neuhaus, J. M. (2011). Misspecifying the Shape of a Random Effects Distribution: Why Getting It Wrong May Not Matter. *Statistical Science*, *26*(3), 388–402.

Merrmann, M., Walsh, E., Isenberg, E., & Resch, A. (2013). *Shrinkage of Value-Added Estimates and Characteristics of Students with Hard-to-Predict Achievement Levels - value-added_shrinkage_wp.pdf* (Working Paper No. 17). Prtinceton, NJ: Mathematica Policy Research. Retrieved from http://www.mathematica-mpr.com/~/media/publications/PDFs/education/value-added_shrinkage_wp.pdf

Mihaly, K., McCaffrey, D., Sass, T. R., & Lockwood, J. R. (2013). Where You Come From or Where You Go? Distinguishing Between School Quality and the Effectiveness of Teacher Preparation Program Graduates. *Education Finance and Policy*, *8*(4), 459–493.

Papay, J. P., & Kraft, M. A. (2015). Productivity returns to experience in the teacher labor market: Methodological challenges and new evidence on long-term career improvement. *Journal of Public Economics*, *130*, 105–119. http://doi.org/10.1016/j.jpubeco.2015.02.008

Pustejovsky, J. E., & Tipton, E. (2016). Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *arXiv:1601.01981 [stat]*. Retrieved from http://arxiv.org/abs/1601.01981

Rothstein, J. (2014). Revisiting the Impacts of Teachers. *Unpublished Manuscript*.

Rumberger, R. W., & Thomas, S. L. (1993). The economic returns to college major, quality and performance: A multilevel analysis of recent graduates. *Economics of Education Review*, *12*(1), 1–19. http://doi.org/10.1016/0272-7757(93)90040-N

Staiger, D. O., & Rockoff, J. E. (2010). Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives*, *24*(3), 97–118. http://doi.org/10.1257/jep.24.3.97

Texas Education Agency. (2011). Technical Digests and Reports, 2010-2011. Retrieved September 27, 2014, from http://www.tea.state.tx.us/student.assessment/techdigest/

Texas State Legislature. Texas Senate Bill 174 (2009).

US Department of Education. (2011). *Our Future, Our Teachers: The Obama Administration's Plan for Teacher Education Reform and Improvement*. Washington DC.

von Hippel, P. T. (2015). The heterogeneity statistic I2 can be biased in small meta-analyses. *BMC Medical Research Methodology*, *15*(1), 35.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical Principles In Experimental Design* (3 edition). New York: McGraw-Hill Humanities/Social Sciences/Languages.

Wiswall, M. (2013). The dynamics of teacher quality. *Journal of Public Economics*, *100*, 61–78. http://doi.org/10.1016/j.jpubeco.2013.01.006

Published in 2016 by *Economics of Education Review* 53:31-45.

Tables and Figures

Table 1. Sample sizes, all grades

	Math	Reading
Students	298,584	210,397
Teachers	6,358	4,965
Classrooms	24,008	17,660
Schools	3,491	3,085
Districts	765	711
TPPs	95	92

Note. The sample is limited to teachers in their 1st, 2nd, or 3rd year of teaching. This is the largest and most diverse sample ever used to estimate teacher quality differences between TPPs.

Table 2. Estimates, SEs, and significance of all-grade TPP estimates

a. All TPPs

Subject	Model	TPPs	SD of point estimates	Mean of SEs	Corr. with OLS	Corr. with RE teachers	TPPs significantly different	
							Uncorrected	Bonferroni corrected
Math	OLS	95	.071	.023	1	.89	43	23
	OLS with teacher clustered SEs		.071	.043	1	.89	16	2
	OLS with school clustered SEs		.071	.043	1	.89	14	2
	OLS with district clustered SEs		.071	.041	1	.89	18	4
	OLS with TPP clustered SEs		.071	.011	1	.89	63	48
	RE teachers		.078	.050	.89	1	10	0
	RE schools + RE teachers		.075	.050	.89	.97	7	0
	FE schools + RE teachers	94	.102	.073	.50	.60	5	1
Reading	OLS	92	.054	.027	1	.97	28	5
	OLS with teacher clustered SEs		.054	.039	1	.97	10	1
	OLS with school clustered SEs		.054	.039	1	.97	10	1
	OLS with district clustered SEs		.054	.037	1	.97	14	3
	OLS with TPP clustered SEs		.054	.011	1	.97	56	37
	RE teachers		.056	.041	.97	1	11	1
	RE schools + RE teachers		.051	.041	.95	.98	7	0
	FE schools + RE teachers	91	.081	.067	.14	.16	8	0

b. Large TPPs (≥40 teachers in subject)

Subject	Model	TPPs	SD of point estimates	Mean of SEs	Corr. with OLS	Corr with RE teachers	TPPs significantly different	
							Uncorrected	Bonferroni corrected
Math	OLS	48	.038	.010	1	.85	27	17
	OLS with teacher clustered SEs		.038	.025	1	.85	9	1
	OLS with school clustered SEs		.038	.026	1	.85	8	0
	OLS with district clustered SEs		.038	.026	1	.85	9	1
	OLS with TPP clustered SEs		.038	.007	1	.85	35	25
	RE teachers		.036	.026	.85	1	7	1
	RE schools + RE teachers		.030	.026	.78	.96	3	0
	FE schools + RE teachers	44	.040	.038	.26	.31	2	1
Reading	OLS	37	.022	.013	1	.84	10	4
	OLS with teacher clustered SEs		.022	.020	1	.84	4	0
	OLS with school clustered SEs		.022	.021	1	.84	4	0
	OLS with district clustered SEs		.022	.020	1	.84	6	2
	OLS with TPP clustered SEs		.022	.006	1	.84	21	13
	RE teachers		.022	.020	.84	1	4	1
	RE schools + RE teachers		.021	.020	.76	.94	3	1
	FE schools + RE teachers	35	.044	.033	.10	.15	5	0

Note. The school FE model produces the least precise estimates. Teacher-, school-, and district-clustered SEs are biased downward, but the bias is negligible in large TPPs. OLS and TPP-clustered SEs are biased downward, and the bias does not improve in large TPPs. The models with teacher REs, or teacher and school REs, produce the best SE estimates overall.

Published in 2016 by *Economics of Education Review* 53:31-45.

Table 3. Estimates of all-grade heterogeneity and reliability, calculated using SEs

a. All TPPs

		Reliability		Heterogeneity SD		Homogeneity tests		
	Model	$\hat{\rho}_H$	$\hat{\rho}_Q$ (95% CI)	$\hat{\tau}_H$	$\hat{\tau}_Q$	*df*	*Q*	*LR*
Math	OLS	.82	.88 (.87,.90)	.06	.07	94	813***	820***
	OLS with teacher clustered SEs	.41	.60 (.49,.68)	.05	.06		232***	
	OLS with school clustered SEs	.45	.62 (.53,.70)	.05	.06		243***	
	OLS with district clustered SEs	.52	.67 (.60,.74)	.05	.06		288***	
	OLS with TPP clustered SEs	.97	.97 (.97,.97)	.07	.07		3,253***	
	RE teachers	.40	.38 (.20,.52)	.05	.05		152***	148***
	RE schools + RE teachers	.38	.23 (.01,.41)	.05	.04		122*	122*
	FE schools + RE teachers	.28	.14 (.00,.33)	.05	.04	93	105	
Reading	OLS	.64	.73 (.67,.78)	.04	.05	91	346***	344***
	OLS with teacher clustered SEs	.24	.39 (.22,.53)	.03	.03		151***	
	OLS with school clustered SEs	.24	.41 (.24,.54)	.03	.03		154***	
	OLS with district clustered SEs	.31	.55 (.44,.65)	.03	.04		204***	
	OLS with TPP clustered SEs	.95	.96 (.96,.97)	.05	.05		2,599***	
	RE teachers	.31	.33 (.13,.48)	.03	.03		136**	134**
	RE schools + RE teachers	.18	.24 (.01,.42)	.02	.03		120*	119*
	FE schools + RE teachers	.06	.24 (.01,.42)	.02	.04	90	118*	

b. Large TPPs (≥40 teachers in subject)

		Reliability		Heterogeneity SD		Homogeneity test	
	Model	$\hat{\rho}_H$	$\hat{\rho}_Q$ (95% CI)	$\hat{\tau}_H$	$\hat{\tau}_Q$	*df*	*Q*
Math	OLS	.92	.92 (.90, .93)	.04	.04	47	573***
	OLS with teacher clustered SEs	.50	.49 (.29, .64)	.03	.03		94***
	OLS with school clustered SEs	.49	.46 (.24, .62)	.03	.03		87***
	OLS with district clustered SEs	.44	.59 (.44, .70)	.03	.03		116***
	OLS with TPP clustered SEs	.95	.96 (.96, .97)	.04	.04		1,315***
	RE teachers	.43	.46 (.24, .62)	.02	.02		87***
	RE schools + RE teachers	.18	.22 (.00, .46)	.01	.01		60
	FE schools + RE teachers	.09	.16 (.00, .42)	.01	.01	43	49
Reading	OLS	.63	.78 (.70, .84)	.02	.02	36	169***
	OLS with teacher clustered SEs	.04	.35 (.03, .57)	.00	.01		55*
	OLS with school clustered SEs	.01	.34 (.00, .56)	.00	.01		53*
	OLS with district clustered SEs	.07	.53 (.32, .68)	.01	.02		76***
	OLS with TPP clustered SEs	.91	.96 (.95, .96)	.02	.02		849***
	RE teachers	.05	.36 (.04, .57)	.01	.01		55*
	RE schools + RE teachers	.00	.29 (.00, .53)	.00	.01		50
	FE schools + RE teachers	.00	.21 (.00, .48)	.00	.01	34	56*

Note. Models which underestimate the SEs will overestimate reliability and heterogeneity; they will also over-reject the null hypothesis of homogeneity. Models that provide better SE estimates suggest that little heterogeneity is present, especially in large TPPs.

Table 4. Estimates of all-grade heterogeneity and reliability, calculated by correlating single-grade point estimates

a. All TPPs

		Reliability	Heterogeneity SD	Homogeneity test
Subject	Model	$\hat{\rho}_{ICC}$ (95% CI)	$\hat{\tau}_{ICC}$ (95% CI)	*F*
Math	OLS (with or without clustered SEs)	.40 (.15,.57)	.03 (.02,.04)	1.68***
	RE teachers	.42 (.18,.58)	.03 (.02,.04)	1.73***
	RE schools + RE teachers	.38 (.12,.55)	.03 (.01,.04)	1.62***
	FE schools + RE teachers	.23 (.00,.44)	.04 (.00,.06)	1.29
Reading	OLS (with or without clustered SEs)	.26 (.00,.47)	.02 (.00,.03)	1.34*
	RE teachers	.36 (.08,.55)	.02 (.01,.03)	1.57**
	RE schools + RE teachers	.31 (.00,.51)	.02 (.00,.03)	1.45*
	FE schools + RE teachers	.00 (.00,.30)	.00 (.00,.05)	0.95

b. Large TPPs (≥40 teachers in subject)

		Reliability	Heterogeneity SD	Homogeneity test
Subject	Model	$\hat{\rho}_{ICC}$ (95% CI)	$\hat{\tau}_{ICC}$ (95% CI)	*F*
Math	OLS (with or without clustered SEs)	.37 (0,.58)	.02 (0,.03)	1.59*
	RE teachers	.37 (0,.58)	.02 (0,.03)	1.60*
	RE schools + RE teachers	.19 (0,.46)	.01 (0,.02)	1.24
	FE schools + RE teachers	.10 (0,.40)	.02 (0,.04)	1.11
Reading	OLS (with or without clustered SEs)	.10 (0,.43)	.01 (0,.02)	1.11
	RE teachers	.11 (0,.44)	.01 (0,.02)	1.13
	RE schools + RE teachers	.10 (0,.43)	.01 (0,.02)	1.11
	FE schools + RE teachers	.34 (0,.60)	.04 (0,.07)	1.51

Note. These reliability and heterogeneity estimates do not depend on SE estimates. Across all models they suggest that little heterogeneity is present.

Published in 2016 by *Economics of Education Review* 53:31-45.

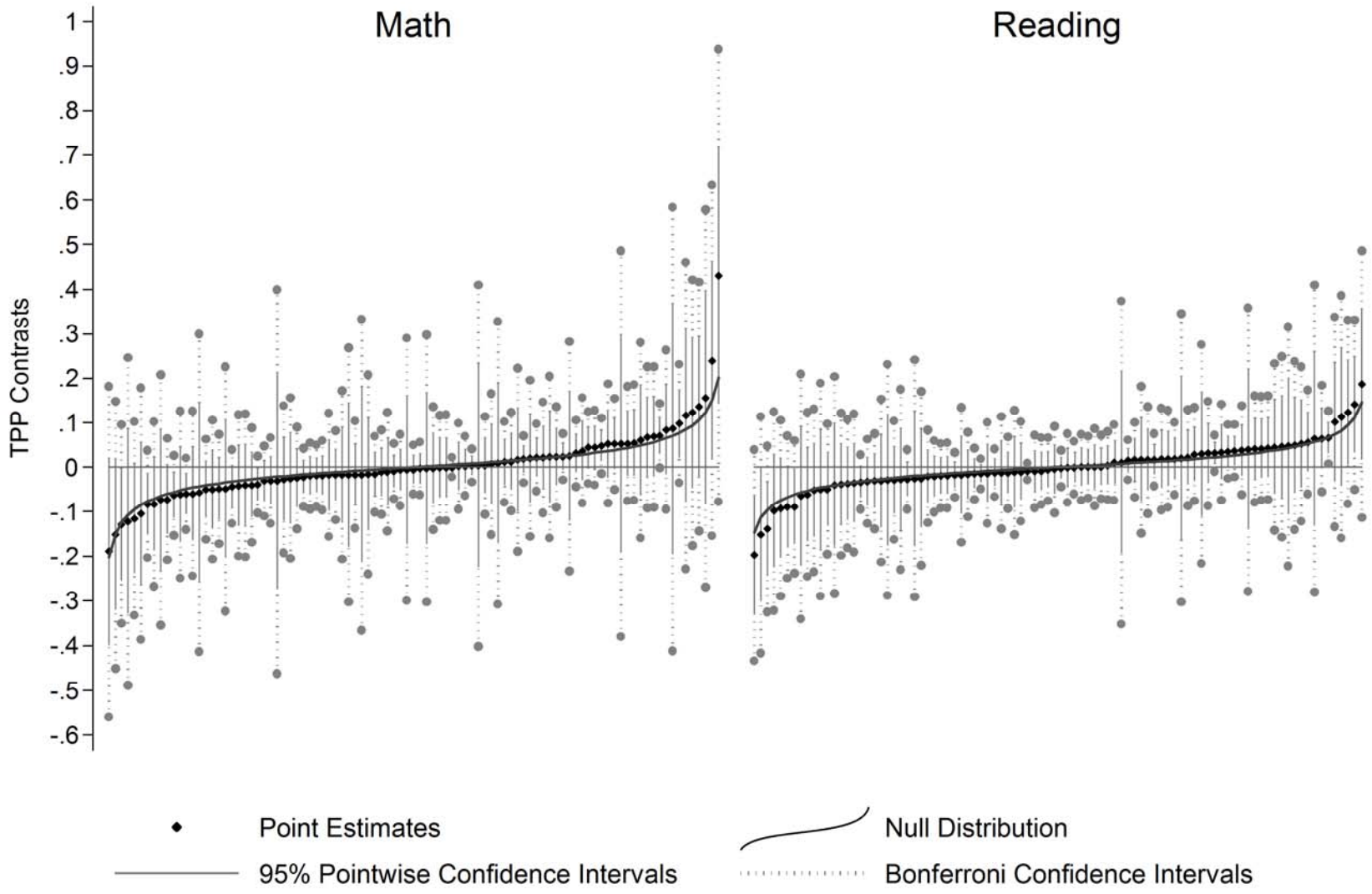

Figure 1. TPP contrasts from the all-grade models. The distribution of point estimates is similar to the null distribution, and only one of the Bonferroni confidence intervals does not cover zero. These results suggest that little true variation between TPPs is present, and most of the observed differences are due to noise.

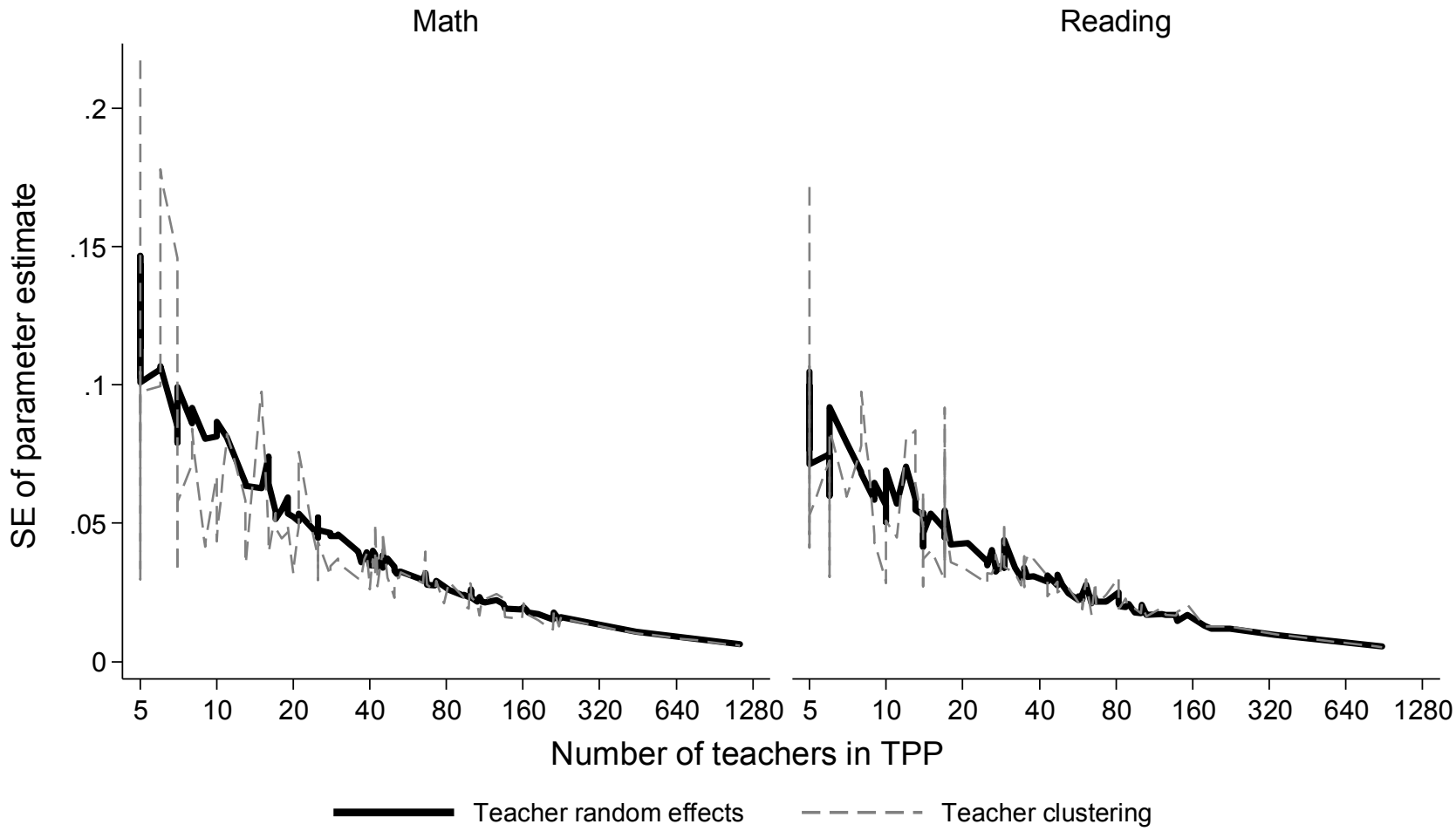

Figure 2. Standard errors (SE) of small and large TPPs under a model with teacher random effects vs. a model with teacher clustered SEs. The teacher-clustered SEs are more volatile and have a negative bias in small TPPs.