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 the quality of their teachers, as estimated by teachers' effects on 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 comparing many TPPs, the multiple comparisons problem increases the danger of classifying ordinary TPPs as good or bad. In addition, there are different statistical methods for comparing TPPs, and the choice of method can affect which TPPs appear to be different. Using Texas data, we compare and evaluate methods for comparing TPPs. We find it is rarely possible to identify TPPs that 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 that prepare teachers for the classroom. While governments have long played a role in approving and funding teacher preparation programs (TPP), sixteen states have begun to practice a new form of TPP accountability that 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 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 does not necessarily follow that all or even most of the good teachers come from good TPPs. To the contrary, it may be that little of the variation in teacher effectiveness lies between TPPs, in which case the differences between teachers from good and bad TPPs may not be large enough for TPP accountability to make a difference.

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. In the extreme, if the estimated differences between TPPs were completely unreliable, 100 percent noise, then shutting down the TPP with the worst estimate would be equivalent to shutting down a TPP at random.

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 (or equivalently, if we calculate multiple 95 percent confidence intervals [CIs]), 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. Even in a state with just 20 TPPs, all identical, there would be a 64 percent chance ($1–.95^{20}$) of erroneously concluding that at least one TPP is significantly different. 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 question are unsettled. 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, 2012). 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 Missouri 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, and how to cluster standard errors (Koedel, Parsons, et al., 2015; Lincove, Osborne, Dillon, & Mills, 2014; Mihaly, McCaffrey, Sass, & Lockwood, 2013), and there are further modeling issues, such as whether to include random effects at the teacher, school, or district levels (e.g., Gansle et al., 2012). Once a model has been fit, the methodological decisions are note 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 standard error (SE) estimates are more sensitive and can be biased and volatile. While the 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 the estimates of those differences consist mostly of noise, even in large TPPs. We also find that hardly any TPPs can be flagged as different after adjustments are made for multiple tests. These results suggest that TPP accountability systems have very limited potential to improve student achievement. In

addition, careless reading 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 the effects of TPPs on 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 US state, Texas offers a lot of 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.

Table 2 lists the TPPs in our data. Texas TPPs are diverse in both size and approach. The largest TPP, at the top of the table, contributed over 1,000 math teachers to our data; the smallest TPPs, near the bottom, contributed only 4 each. Although many Texas TPPs are traditional programs run out of colleges and universities, the state's four largest TPPs are newer "alternative" TPPs, three of which are run for profit. Other TPPs are run by independent school districts (ISDs) and regional educational service centers (ESCs) 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 grade, 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 and comparability.[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. Teachers who were not certified in math or reading 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 ones. 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 that students 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 teachers' control. To capture those influences, we coded a number of classroom, school, and district variables. At the classroom level, we calculated the class size (number of students) as well as 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 include 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 include 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 is larger in low-income districts (Corcoran, Evans, Godwin, Murray, & Schwab, 2004). We also include 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

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

and our model for value added to 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. Notice that we use lagged scores from two different subjects, which reduces 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 raises missing-data issues 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 fourth-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. An alternative to using school and district covariates is to use school fixed effects. But school fixed effects would reduce the analytic sample to 1st-3rd year teachers who work in the same schools as 1st-3rd year teachers from other TPPs. The reduction in the analytic

sample yields larger SEs, and potential for bias if the teachers and schools in the limited sample are not representative of the larger population (Mihaly et al., 2013).

$\boldsymbol{TPP}_t$ 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.

3.2 Estimates and contrasts

From the TPP model, we get estimated TPP coefficients $\hat{\alpha} = [\hat{\alpha}_1 \quad \hat{\alpha}_2 \quad \ldots \quad \hat{\alpha}_P]$ as well as *contrasts* $\Delta\alpha_p = \alpha_p - E(\alpha)$ which are defined as the difference between the p^{th} TPP coefficient α_p and the average coefficient $E(\alpha)$. $E(\alpha)$ can be estimated by the simple mean $\bar{\alpha}$, the student weighted mean $\bar{\alpha}_n$, or the precision-weighted mean $\bar{\alpha}_{s2}$:

$$\begin{aligned} \bar{\alpha} &= \frac{1}{P}\sum_{p=1}^{P}\hat{\alpha}_p \\ \bar{\alpha}_n &= \frac{\sum n_p\hat{\alpha}_p}{\sum n_p} \\ \bar{\alpha}_{s2} &= \frac{\sum s_p^{-2}\hat{\alpha}_p}{\sum s_p^{-2}} \end{aligned} \tag{2}$$

where n_p is the number of students taught by teachers from the p^{th} TPP, and the precision s_p^{-2} is the inverse of the squared SE. All three estimators are consistent, but $\bar{\alpha}_{s2}$ is the most efficient, and $\bar{\alpha}_n$ is similar. We will use the symbols $\Delta\hat{\alpha}$, $\Delta\hat{\alpha}_n$, $\Delta\hat{\alpha}_{s2}$, respectively, for contrasts that use the simple mean, the student-weighted mean, or the precision-weighted mean.

The TPP estimates $\hat{\alpha}_p$ have a covariance matrix V whose diagonal terms are the squared standard errors $s_p^2 = SE^2(\hat{\alpha}_p)$, $p = 1,\ldots,P$.[2] In large samples, the covariance matrix is practically the same for the estimates $\hat{\alpha}$ as for the contrasts $\Delta\hat{\alpha}$, $\Delta\hat{\alpha}_n$, or $\Delta\hat{\alpha}_{s2}$.

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.

3.3 Clustering

An important issue is that the residuals e_i in equation (1) are correlated among students who are taught by the same teacher. One way to account for within-teacher correlation is to estimate clustered SEs. Clustered SEs work by calculating residuals around the OLS estimates, estimating the within-cluster covariance matrix of the residuals, and using that matrix to estimate the SE. Past TPP research has recommended clustering at the teacher level (Koedel, Parsons, et al., 2015), but it is also plausible to cluster at a higher level 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 clustering 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*. 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 available 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; McCaffrey, personal communication, June 10, 2015).

3.4 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 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.

Above the teacher level, we can add school or district REs to get a multilevel or hierarchical linear model (HLM) with REs at two or three nested levels (Raudenbush & Bryk, 2001). An HLM with teacher and school REs has been used to estimate TPP coefficients in Louisiana (Gansle et al., 2012). The decision of whether to add a level of REs can be made with the likelihood ratio test

$$LR_{RE} = 2(\ell_1 - \ell_0) \qquad (3)$$

where ℓ_1 and ℓ_0 are the likelihoods with and without the added level of REs. For example, the LR_{RE} test can be used to choose between an OLS model and a model with teacher REs, or between a model with teacher REs and a model with teacher and school REs, or between a model with teacher and school REs vs. a model with teacher, school, and district REs.

Because REs cannot have negative variance, the LR_{RE} test is one-sided (e.g., $H_1: \sigma_r^2 > 0$ vs $H_0: \sigma_r^2 = 0$). To get the *p* value for an LR_{RE} test, we first calculate a *p* value from a χ_1^2 distribution, and then cut that *p* value in half (LaHuis & Ferguson, 2009; Stram & Lee, 1994).[3]

3.5 Multiple comparisons

It is common to plot all the TPP contrasts $\Delta\hat{\alpha}_p$ with ordinary 95% pointwise CIs. 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 testing at a .05 significance level each of the P hypotheses $H_0: \Delta\alpha_p = 0$, p=1,…,P.

The problem with this approach is that it makes 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), we would expect to conclude that approximately five differ significantly from the average—even if all are in fact identical.[4]

The simplest adjustment for multiple comparisons is the Bonferroni correction. Under the Bonferroni correction, we test at a significance level of .05/P or, equivalently, construct 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.

3.6 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

$$V(\hat{\alpha}_p) = \tau^2 + \sigma^2 \quad (4)$$

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. Then the fraction of variance in $\hat{\alpha}_p$ that is due to heterogeneity rather than error is

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2} \quad (5).$$

We call ρ the *reliability* of the estimates $\hat{\alpha}_p$. 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:

$$\begin{aligned} H_0&: \alpha_1 = \alpha_2 = \cdots = \alpha_P \\ \text{or } H_0&: \tau^2 = 0 \\ \text{or } H_0&: \rho = 0 \end{aligned} \quad (6)$$

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[5] normal distributions with means of zero and different variances. We can approximate this mixture using a simple procedure described in this footnote.[6] Under H_0, the TPP contrasts $\Delta\hat{\alpha}_p$ would approximate the (P+1)-quantiles from $\mathcal{D}_0$ (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 evidence that signal is present and may be able to highlight specific TPPs as exceptional.

3.7 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?

3.7.1 Using TPP point estimates

The simplest approach is to compare different point estimates of the same TPP coefficients. From our models we get TPP point estimates for different subjects (reading, math), and for different grades (4^{th}-9^{th} in reading, 4^{th}-10^{th} in math). We could also get TPP point estimates for different cohorts of teachers. In different data, we could get estimates for different school years.

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 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

$$\Delta\hat{\alpha}_{pj} = \Delta\alpha_p + u_{pj} \qquad (7),$$

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, which we call F_{ICC}. Standard ANOVA formulas[7] (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 grades—then the reliability of the average is estimated by (Winer, Brown, & Michels, 1991).

$$\hat{\rho}_{\text{ICC}} = \frac{Jr}{1 + (J-1)r} \tag{8}$$

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. For example, independence is violated if we have TPP estimates for two different school years, but many of the teachers are the same in both years.

If estimates are independent but not exchangeable, the consequences are less dire. Reliability will be underestimated, but we can still reject the null hypothesis if the estimated reliability is significantly greater than zero. For example, reading and math estimates 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. In that case, the correlation between reading and math estimates will be lower than the reliability of either set of estimates considered by itself, and the covariance between the reading and math estimates will lie somewhere between the heterogeneity of the reading estimates and the heterogeneity of the math estimates.

The assumptions of independence and exchangeability are more plausible when we are comparing estimates for the same subject in nearby grades. For example, 4$^{\text{th}}$ and 5$^{\text{th}}$ grade math teachers from the same TPP are both independent and exchangeable. 4$^{\text{th}}$ and 10$^{\text{th}}$ grade math teachers from the same TPP are also independent, but may not be exchangeable if they are trained differently or if different skills are needed to teach math in 4$^{\text{th}}$ vs. 10$^{\text{th}}$ grade.

3.7.2 Using SEs

If we need to assess the heterogeneity and reliability of a single set of TPP estimates, then the point estimates by themselves are not sufficient. We also need the SEs or a related statistic, such as the log likelihood. This makes our estimates of heterogeneity and reliability somewhat sensitive since, as we discussed earlier, the SEs and likelihood can be estimated in different ways and from different models.

We use several statistics to test the null hypothesis of homogeneity. The likelihood ratio statistic LR_{TPP} compares the log-likelihoods ℓ of the TPP and no-TPP models:

$$LR_{TPP} = 2(\ell_{TPP} - \ell_{noTPP}) \quad (9)$$

The Wald statistic W compares the contrasts to their estimated covariance matrix $\hat{V}$:

$$W = \Delta\hat{\alpha}_n \hat{V}^{-1} \Delta\hat{\alpha}_n^T \quad (10),$$

The Cochran statistic Q compares the contrasts to their estimated standard errors, whose squares $\hat{s}_p^2$ are on the diagonal of $\hat{V}$:

$$\begin{aligned} Q &= \Delta\hat{\alpha}_{s2} \left(diag\left(\hat{V}\right) \right)^{-1} \Delta\hat{\alpha}_{s2}^{\mathrm{T}} \\ &= \sum_{p=1}^{P} \frac{\Delta\hat{\alpha}_{s2,p}^2}{\hat{s}_p^2} \end{aligned} \quad (11)$$

Under the null hypothesis of homogeneity, LR_{TPP}, W, and Q all follow a χ^2_{P-1} distribution if the sample is large and the model is correctly specified. Q has long been used in meta-analysis (Cochran, 1954), and was recently introduced to the teacher and TPP literatures (Koedel, 2009; Koedel, Parsons, et al., 2015). W has also been used in the TPP literature (Koedel, 2009; Koedel, Parsons, et al., 2015),[8] as has its transformation, the regression F statistic $F_{reg} = W/(P-1)$ (Goldhaber et al., 2013).

Q is the most convenient statistic. It uses a scalar formula, widely implemented in meta-analysis software, which can be easily calculated from a regression table reporting TPP estimates and SEs. Q is

suitable for secondary analysis of TPP estimates reported by others, and it can be applied to a subset of TPP estimates—e.g., estimates for only the largest TPPs.

W is a little less convenient. It is a matrix formula that requires the off-diagonal elements of $\hat{V}$, so it cannot be calculated from the estimates and SEs in a regression table. But *W* is often provided conveniently by software such as the *contrast* postestimation command in Stata.

The LR_{TPP} statistic is the least convenient. One inconvenience is that LR_{TPP} cannot be used with clustered SEs, because the calculation of the likelihood ignores the clustering. Another inconvenience is that, unlike the *W* and *Q* statistics, the LR_{TPP} statistic applies to the whole model and cannot be calculated from a subset of estimates—e.g., only the estimates for large TPPs. If you want to calculate LR_{TPP} for a subset of TPPs, you have to limit the data to those TPPs and refit both the TPP model and the no-TPP models to get their respective likelihoods.

The *Q* statistic can be transformed into an estimate of reliability:

$$\hat{\rho}_Q = \max\left(0, 1 - \frac{P-1}{Q}\right) \tag{12}$$

This reliability estimate is widely used in meta-analysis, where is it called I^2 and comes with a test-based CI (Higgins & Thompson, 2002; von Hippel, 2015). It was recently introduced to the teacher and TPP literatures (Koedel, 2009; Koedel, Parsons, et al., 2015).

The heterogeneity variance can be estimated as the difference between the variance of the estimates and the variance of the null distribution.

$$\begin{aligned} \hat{\tau}_H^2 &= \max\left(0, \hat{V}(\hat{\alpha}_p) - \hat{V}(\mathcal{D}_0)\right) \\ \text{where } \hat{V}(\mathcal{D}_0) &= \frac{1}{P}\sum \hat{s}_p^2 \\ \text{and } \hat{V}(\hat{\alpha}_p) &= \frac{1}{P-1}\sum\left(\hat{\alpha}_p - \overline{\alpha}\right)^2 \end{aligned} \tag{13}$$

This estimator has long been used in meta-analysis (Hedges, 1983), and was recently introduced to the teacher and TPP literatures (Aaronson, Barrow, & Sander, 2007; Koedel, 2009; Koedel, Parsons, et al., 2015).

Another heterogeneity statistic that is used in meta-analysis is (DerSimonian & Laird, 1986)

$$\hat{\tau}^2_{DL} = \max\left(0, \frac{Q-(P-1)}{\sum \hat{s}_p^{-2} - \sum \hat{s}_p^{-4} / \sum \hat{s}_p^{-2}}\right) \tag{14}$$

In the TPP literature, another estimate of the heterogeneity variance is $\hat{\tau}^2_{EB}$, which is the variance of the empirical Bayes (EB) contrasts $\Delta\tilde{\alpha}_p$ (Boyd et al., 2009; Goldhaber et al., 2013). The EB contrasts shrink the contrasts $\Delta\hat{\alpha}_p$ by an estimate $\hat{\rho}_p$ of their reliability (Merrmann, Walsh, Isenberg, & Resch, 2013):

$$\begin{aligned} \Delta\tilde{\alpha}_p &= \hat{\rho}_p \Delta\hat{\alpha}_p \\ \text{where } \hat{\rho}_p &= \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{s}_p^2} \end{aligned} \tag{15}$$

Note that the shrinkage factor $\hat{\rho}_p$ estimates the reliability of the individual contrast $\Delta\hat{\alpha}_p$, which is different from the overall reliability ρ of all the contrasts together. Note also that the calculation of $\hat{\rho}_p$ requires some prior estimate $\hat{\tau}^2$ of the heterogeneity variance; we use $\hat{\tau}^2_{DL}$.

Any estimate of heterogeneity τ^2 implies an estimate of reliability ρ, and vice versa, through the relationship $\rho = \tau^2 / \hat{V}(\hat{\alpha}_p)$. For example, the reliability estimate $\hat{\rho}_Q$ implies the following heterogeneity estimate: $\hat{\tau}^2_Q = \hat{\rho}_Q \hat{V}(\hat{\alpha}_p)$. Likewise the heterogeneity estimate $\hat{\tau}^2_H$ implies a reliability estimate $\hat{\rho}_H = \hat{\tau}^2_H / \hat{V}(\hat{\alpha}_p)$. The estimates $\hat{\tau}^2_Q$ and $\hat{\rho}_H$ have been used in the teacher and TPP literatures (Koedel, 2009; Koedel, Parsons, et al., 2015).

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. Caterpillar plots are a common way to present estimates, and at first glance Figure 1 might seem to offer a lot of information about the differences between TPPs. But in fact the differences between TPP estimates are mostly noise.

Figure 1 demonstrates the noisiness of the estimates explicitly 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. In Figure 1, the null distribution is only a little less dispersed than the observed TPP distribution, suggesting that the observed distribution contains a lot of noise and only a 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 if the null distribution, which assumes no TPP differences, has a similar 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. The null distribution is uncannily similar to the observed distribution of TPP estimates. Only on the far right of the math caterpillar plot, and only in the extremes of the reading caterpillar plot, are the observed TPP estimates noticeably more dispersed than the null distribution.

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 the reference

line, 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 95 comparisons in math and 92 comparisons 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 better than average in reading. Note that the significant reading estimate does not have the largest point estimate in the caterpillar plot. Instead, it has the 6th largest positive point estimate and the 12th smallest SE.

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. a summarizes the distribution of TPP point estimates, SE estimates, and CIs under different OLS and RE models.

The point estimates are fairly robust to model choice. According to Table 3a, the correlation between OLS and RE point estimates is .88-.89 in math and .95-.97 in reading. The standard deviation (SD) of the TPP point estimates is also very similar, regardless of which model is chosen.

The estimated SEs are more sensitive to model specification. At the top of Table 3a, we have OLS, which substantially underestimates the SEs because it fails to account for residual correlation, especially at the teacher level. Lower in the table, we account for teacher-level correlation by using teacher clustering or teacher REs. Both options substantially increase the SE estimates. Teacher-clustered SEs are 87% larger than OLS SEs in math and 44% larger than OLS SEs in reading. The SEs from a teacher RE model are even larger—117% larger than OLS SEs in math, and 52% larger than OLS SEs in reading.

Why are estimated SEs larger with teacher REs than with teacher clustering? The reason is that teacher-clustered SEs are negatively biased for the half of TPPs that have fewer than 40 teachers (Cameron & Miller, 2015). For these small TPPs, clustered SEs are not just biased, but also volatile (Bell & McCaffrey, 2002). Figure 2 shows how SE estimates change with TPP size. With teacher REs, SE estimates decrease smoothly with the inverse square root of the number of teachers. But teacher-clustered SEs, in addition to being too small on average, are extremely volatile and do not decrease smoothly with sample size until the number of teachers goes above 40. With more than 40 teachers, the differences between teacher-clustered SEs and SEs from a teacher RE model become inconsequential.

Clustering at higher levels does not improve the negative bias of clustered SEs; in fact, clustering at higher levels can worsen the bias by reducing the number of clusters. This is evident in Table 3a. With TPP clustering there is only one cluster per TPP, and the resulting SE estimates are so negatively biased that they are actually smaller than OLS SEs. With school clustering, SEs are about the same bias as they have with teacher clustering, because a small TPP will typically place each of its teachers in a different school. District-clustered SEs are a little smaller, with a little more negative bias than teacher- or school-clustered TPPs, because it is not uncommon for a TPP to place several teachers in the same large district.

Table 3a shows that adding REs at the school and district levels has little effect on the SE estimates, but slightly shrinks the dispersion of the point estimates. This is probably because RE point estimates are a compromise between OLS and fixed effects (FE) estimates, so that RE estimates remove from the point estimates some of the between-school and between-district variation that is not explained by covariates (Greene, 2011; Wooldridge, 2001).

Even small differences between estimates from different models can affect our conclusions about which TPPs are different. We will of course get too many spuriously significant differences if we

neglect the Bonferroni correction, or if we fit a model that yields very biased SEs (e.g., OLS or TPP clustering). But even Bonferroni intervals from plausible clustered or RE models can differ in their implications. For example, Bonferroni tests using clustered SEs at the teacher, school, or district levels suggest that 2-4 TPPs are significantly different from average in reading and 1-3 are significantly different in math. But Bonferroni tests from RE models suggest that zero TPPs are significantly different in reading and 0-1 are significantly different in math.

Given the sensitivity of TPP estimates to model specification, it would be helpful to have evidence regarding which models are, in some sense, better. Figure 2 provided some evidence by reminding us that teacher-clustered SEs are biased and volatile in smaller TPPs. This is one reason to favor an RE model if a state plans to include smaller TPPs in its accountability system.

If an RE model is chosen, then Table 4 suggests that it is better to use a model with school and possibly district REs as well as teacher REs. Each level of RE has a SD that is significantly greater than zero, and LR_{RE} tests show that each level of RE significantly improves the fit of the model. A model with teacher REs fits significantly better than an OLS model; a model with teacher and school REs fits significantly better than a model with teacher REs alone; and a model with teacher, school, and district REs fits significantly better than a model with only teacher and school REs. The school REs are comparable in size to the teacher REs, but the district REs are considerably smaller. Models with school and district REs yield larger and more accurate SE estimates for school and district covariates, and may affect TPP estimates when there are multiple teachers in the same school or districts. a, however, suggests that the -school and district REs change the TPP estimates very little.

(The model chosen has implications not just for the TPP estimates, but for the time that it takes to obtain them. Applying Stata to our data, we can obtain OLS point estimates in 1-2 minutes, but clustered

SEs take about half an hour, and RE estimates take 3-11 hours. Evaluators who favor RE models should consider HLM software, which fits RE models to large datasets much more quickly than Stata or SAS.)

4.3 Limiting accountability to large TPPs

Both policy and statistical arguments can be made for limiting accountability to large TPPs. One statistical argument is that it is difficult to estimate the coefficients of small TPPs precisely, and some estimation methods, such as clustered SEs, are biased and volatile in small TPPs. If clustered SEs are used, a plausible policy is to limit accountability to TPPs with at least 40 teachers; if RE models are used, then accountability may be extended to somewhat smaller TPPs. Another statistical argument for limiting accountability to large TPPs is that it reduces the number of TPPs that must be compared, and thatat reduces the multiple comparisons problem. It is still necessary to correct for multiple comparisons, but the corrected CIs will be narrower if there are fewer comparisons to correct for.

A policy argument for limiting accountability to larger TPPs is simply that larger TPPs affect more students. For example, if two TPPs, one large and one small, are certifying equally poor teachers, we can have greater impact by shutting down the large TPP than by shutting down the small one. Conversely, if a large and small TPP are certifying equally *good* teachers, then the larger TPP may find it easier to expand.

One argument against focusing on large TPPs is that smaller TPPs may be more heterogeneous. If smaller TPPs are more heterogeneous, then an accountability system that looks for large TPPs whose teachers are slightly above average may miss a small TPP whose teachers are truly extraordinary. On a per-child basis, a small number of extraordinary teachers may have impact as great as a large number of teachers who are barely above average.

Table 3b summarizes the point estimates, SEs, and CIs 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.

Compared to the SEs for small TPPs, the SE estimates for large TPPs are smaller and less sensitive to model specification. The OLS and TPP-clustered SEs still have a severe negative bias, but the SEs from other models are approximately unbiased and agree closely with one another. This is partly because clustered SEs have little bias when the number of clusters is large.

Despite the smaller, less sensitive SEs of large TPPs, and despite the fact that a sample limited to large TPPs has fewer comparisons to correct for, it remains difficult to single out specific TPPs as significantly different from average. If we limit our attention to plausible models (excluding OLS and TPP clustering), we find that Bonferroni tests flag few if any large TPPs as significantly different. Under some model specifications, no large TPPs are significantly different, and under other specifications only 1-2 large TPPs are significantly different.

A possible reason for the rarity of significantly differences among large TPPs is that the coefficients of large TPPs, though more precisely estimated, may be less heterogeneous. We will find some evidence that large TPPs are less heterogeneous in the next section.

4.4 Heterogeneity and reliability

In light of our difficulty highlighting individual TPPs that are significantly different, one might begin to doubt that there are any teacher quality differences between TPPs at all. In fact, there are differences, but they are very small, especially among large TPPs, and they are not very reliably estimated. We estimate heterogeneity and reliability in this section.

4.4.1 Using point estimates

The most robust way to estimate reliability and heterogeneity is to look at the correlation and covariance between TPP point estimates. As we have seen, TPP point estimates are less sensitive to model specification than are SE estimates. If we get point estimates using OLS, the point estimates do not change if we cluster the SEs. If we get point estimates using an RE model, the point estimates are strongly correlated with the OLS point estimates.

Table 5a gives the correlation between the reading and math point estimates for all 87 TPPs that have estimates in both subjects. The correlation is statistically significant and about 0.4, indicating that the estimates are about 40% reliable. Table 5a also reports the square root of the covariance, which is an estimate the heterogeneity SD. The estimated heterogeneity SD is 0.04, which indicates that a 1 SD increase in TPP quality predicts a 0.04 SD increase in student test scores. These estimates of reliability and heterogeneity change very little with model specification.

Table 5b gives the same statistics for large TPPs. One might expect large TPP estimates to be more reliable, but they are not; the correlation between reading and math estimates is no stronger for large TPPs than it is for all TPPs together. This is because the heterogeneity SD, which was 0.04 for all TPPs, is just 0.015-0.019 for large TPPs. In other words, although estimates for large TPPs have less noise, they also have less signal, so on balance the ratio of signal to noise is no better for large TPPs than it is for other TPPs.

Instead of comparing estimates across subjects, we can compare estimates for the same subject across different grades. Table 6a gives the resulting estimates of reliability and heterogeneity for all TPPs. In math, the estimates, averaged across grades 4-10, are about 40% reliable and have a small but significant heterogeneity SD of 0.03. In reading, the estimates across grades 3-9 are about 30% reliable and have a small but significant heterogeneity SD of 0.02. Again, these estimates change little with model specification.

Table 6b gives the same estimates for large TPPs. Again, the results suggest that there is less heterogeneity among large TPPs than among small ones. The all-grade estimates are just 14-37% reliable. They have an estimated heterogeneity SD of just 0.01-0.02, and under two model specifications, we cannot reject the hypothesis that there is no heterogeneity at all. In reading, reliability is estimated at about 10%, the heterogeneity SD is estimated at 0.01, and there is no model specification where we can reject the null hypothesis that there is no heterogeneity at all.

4.4.2 Using SE estimates

Instead of comparing different sets of TPP point estimates, we can test for heterogeneity by comparing a single set of point estimates to their SEs (the *Q* statistic) or to their covariance matrix (the *W* statistic). Or we can calculate the increase in log likelihood when TPP indicators are added to the model (the LR_{TPP} statistic).

Table 7 displays the LR_{TPP}, *Q* and *W* tests of homogeneity among all TPPs, as well as the *Q* test for large TPPs. Among all TPPs, the tests reject the null hypothesis of homogeneity except when there are district REs. Among large TPPs, the *Q* test rejects the null hypothesis except when there are school REs or school and district REs. The LR_{TPP} and *Q* statistics are practically identical, but the LR_{TPP} statistic cannot be used with clustering. The *W* statistic is similar except when there is clustering at the district or TPP level. With TPP clustering, the value of *W* seems implausibly large. Since the only material difference between the *Q* and *W* formulas is that *W* uses the off-diagonal elements of the covariance matrix $\hat{V}$, those off-diagonal elements must account for the large *W* values with district or TPP clustering. We speculate that the off-diagonal elements are very imprecisely estimated.

All of the test statistics in Table 7 are sensitive to model specification, returning values that are too large when the estimated SEs are too small. In particular, if we use OLS or TPP clustering, the SEs are substantially underestimated, so the test statistics are much too large. Similarly, if we use teacher,

school, or district clustering, the SEs are slightly underestimated in small TPPs, so the test statistics are a bit too large unless we limit the estimates to large TPPs. The sensitivity of the statistics in Table 7 contrasts with the robustness of the statistics in Table 5 and Table 6, where test statistics were approximately the same regardless of model specification. The robustness of the statistics in Table 5 and Table 6 results from the fact that they use only the point estimates, ignoring the more sensitive SEs and likelihood.

Table 8 gives estimates of reliability and heterogeneity obtained from estimators that compare the SEs to the point estimates. These estimates are quite sensitive to model specification, returning values that are too large when the estimated SEs are too small. In addition, even when they use the same SEs, the estimators often disagree; in particular, the estimators subscripted with *H* and *Q* return larger estimates of heterogeneity and reliability than the estimators subscripted with *DL* and *EB*.[9] Disagreement among plausible heterogeneity estimators is common (Langan, Higgins, & Simmonds, 2015) and increases uncertainty about how much heterogeneity is actually present. However, all of the estimators agree that the extent of heterogeneity is small.

If we focus on the RE models, which previous results suggest provide the best fit and the least biased SEs, the results suggest that TPP estimates are 0-50 percent reliable, and have a heterogeneity SD of 0-0.05 in math and 0-0.03 in reading. The estimates for large TPPs are less heterogeneous, and no more reliable, than the estimates for all TPPs. These conclusions are broadly compatible with the conclusions that we reached by comparing point estimates in Table 5 and Table 6.

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 consequential.
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 0.03 in math and 0.02 in reading. That is, a 1 SD increase in TPP quality predicts just a 0.03 SD increase in student math scores and a 0.02 SD increase in student reading scores. Under some plausible methods, the heterogeneity SD is as large as 0.05 in math and 0.03 in reading, but under other methods the heterogeneity SD is indistinguishable from 0.

Teacher quality differences between TPPs are not large. 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. The estimates were noisy 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 our TPP estimates would grow more reliable if we had more than one year of data. In addition, having multiple years of data would allow us to estimate reliability not just between subjects and grades but between years as well. To ensure that estimates from different years were independent, we would have to subset the data to ensure that we were looking at different teachers in each year.

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 can get from a single year in Texas.

Uncertainty about TPP estimates is substantial. 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 differed with respect to decisions about clustering and REs, and we have shown that those decisions have some consequences for TPP point estimates and SEs, especially in small TPPs. TPP estimates change even more if a model includes school FEs (Mihaly et al., 2013). In addition, different evaluations have used different sets of covariates, and the selection of covariates can change the distribution of TPP estimates (Lincove et al., 2014).

Finally, it is possible that TPP estimates are not just unreliable but also slightly biased. Possible sources of bias include model misspecification and nonrandom assignment of students to teachers from different TPPs. While 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), the differences between

individual teachers are substantial. Biases that seem small when compared to teacher effects may seem larger when compared to the small differences between TPPs.

Question 3. Which TPPs are different? Even if we are willing to accept the estimates from a single model, it remains hard to single out the specific TPPs that are different. 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 large TPPs, at least in Texas, are very similar in teacher quality.

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 fine 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. It is a little startling that these differences in selectivity and training don't produce bigger differences in teacher effectiveness.

Yet results like this are not unusual in education. 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 students from different schools (Coleman et al., 1966). After college, for 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). 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 0.09 to 0.16 SD in student test scores (Staiger & Rockoff, 2010), it stands to reason that the heterogeneity between TPPs would be as small as 0.01 to .03 SD.[10]

That said, our results are limited to Texas reading and math tests in 2011, and it is possible that results for other years, states, and tests would be different. Results from Missouri and Washington state are similar (Goldhaber et al., 2013; Koedel, Parsons, et al., 2015), but results from New York City and Louisiana suggest larger differences between TPPs (Boyd et al., 2009; Gansle et al., 2012). At the moment it is not clear why estimated TPP difference are larger in some studies than in others. The reasons could be substantive or methodological, and until they are sorted out, there will be some uncertainty regarding the potential of TPP accountability to raise test scores in different states.

It is also possible that results would be different for different outcome variables. Most TPP evaluations have focused on exclusively on reading and math scores, although the Louisiana evaluation also looked at science and social studies scores. It would be informative, though challenging, to estimate between-TPP differences 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 fairly robust to modeling decisions, 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 pointwise 95 percent 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 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 accounting for multiple comparisons—we can have confidence that one TPP is better or worse than average.

We hope that our statistical recommendations will be widely adopted in TPP evaluations and guide policy makers toward careful policy decisions based on a qualified reading of TPP estimates.

5.4 Policy risks and benefits

Even if our recommendations are followed carefully, the benefits of a TPP accountability system would probably be small. Not only are the differences between TPPs small and hard to estimate, but once we identify an exceptional TPP there is no guarantee that we can engineer an effective policy response. Even if an accountability system finds an excellent TPP, there is no guarantee that the TPP can expand without diluting quality. Likewise, even if an accountability system shuts down a poor TPP, there is no guarantee that the TPP that replaces it will be much better. For example, if a TPP is limited by the quality of the local applicant pool, then another TPP screening the same local applicants may not get better results.

Even if evaluators follow our recommendations, we remain concerned that TPP estimates will not be used carefully. Experience has repeatedly shown that, once information is released to leaders and the public, it takes on a life of its own. Asterisks and footnotes are dropped, signal is mistaken for noise, and both leaders and the general public can be "fooled by randomness" (Taleb, 2005). There is a danger that leaders may build TPP accountability on an unstable foundation, basing overconfident actions on evidence that is far less reliable than they imagine.

Endnotes

[1] We excluded students who in 2011 took the Spanish-language TAKS or the "accommodated" TAKS for special education students. We included the scores of students who took the regular TAKS in 2011 but had taken the Spanish or accommodated TAKS in previous years.

[2] We obtained both $\Delta\hat{\alpha}$ and $\hat{V}$ using Stata's postestimation command *contrast gw.TPP*.

[3] An equivalent but more confusing way to say this is that the asymptotic distribution of the LR_{RE} statistic is a 50:50 mixture of χ_1^2 and χ_0^2 distributions, where the χ_0^2 distribution is a point mass at 0 (LaHuis & Ferguson, 2009; Stram & Lee, 1994).

[4] In general, correcting for multiple comparisons means using longer CIs with higher coverage than 95 percent. No past evaluation has done this; in fact, one TPP evaluation corrected in the opposite direction by using 68 percent CIs, which are approximately half as long as 95 percent CIs (Gansle, Noell, & Burns, 2012). The use of 68 percent CIs exacerbates the problem of multiple comparisons. 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.

[5] 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.

[6] We approximate the null distribution using the following procedure. For the p^{th} TPP, the null distribution is $N\left(0, \hat{s}_p^2\right)$, 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., $\widehat{\mathcal{D}}_0 \approx \{q_{1,1}, \ldots, q_{99,1}, q_{1,2}, \ldots, q_{99,2}, \ldots q_{1,P}, \ldots, q_{99,P}\}$. 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.

[7] All these ANOVA calculations are implemented by the *loneway* command in Stata.

[8] Koedel's versions of the W and Q statistic use $\bar{\alpha}$ instead of $\bar{\alpha}_n$ and $\bar{\alpha}_{s2}$ (Koedel, 2009; Koedel, Parsons, Podgursky, & Ehlert, 2015).

[9] The agreement between the DL and EB estimators is not surprising, since in equation (15) we defined the EB estimator as a function of the DL estimator. If we had defined the EB estimator as a function of a different estimator, we might have gotten different results.

[10] 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 $.16 \times \sqrt{.03} \approx .03$.

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.

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.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*(3), 177–188.

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

Greene, W. H. (2011). *Econometric Analysis* (7th ed.). Prentice Hall.

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

Hsu, J. (1996). *Multiple Comparisons: Theory and Methods* (1st ed.). Chapman and Hall/CRC.

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

LaHuis, D. M., & Ferguson, M. W. (2009). The Accuracy of Significance Tests for Slope Variance Components in Multilevel Random Coefficient Models. *Organizational Research Methods*, *12*(3), 418–435. http://doi.org/10.1177/1094428107308984

Langan, D., Higgins, J. P. T., & Simmonds, M. (2015). An empirical comparison of heterogeneity variance estimators in 12 894 meta-analyses. *Research Synthesis Methods*, *6*(2), 195–205. http://doi.org/10.1002/jrsm.1140

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

Raudenbush, S. W., & Bryk, A. S. (2001). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). Sage Publications, Inc.

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

Stram, D. O., & Lee, J. W. (1994). Variance Components Testing in the Longitudinal Mixed Effects Model. *Biometrics*, *50*(4), 1171–1177. http://doi.org/10.2307/2533455

Taleb, N. N. (2005). *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets* (Updated edition). New York: Random House Trade Paperbacks.

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

Wooldridge, J. M. (2001). *Econometric Analysis of Cross Section and Panel Data* (1st ed.). The MIT Press.

Figures

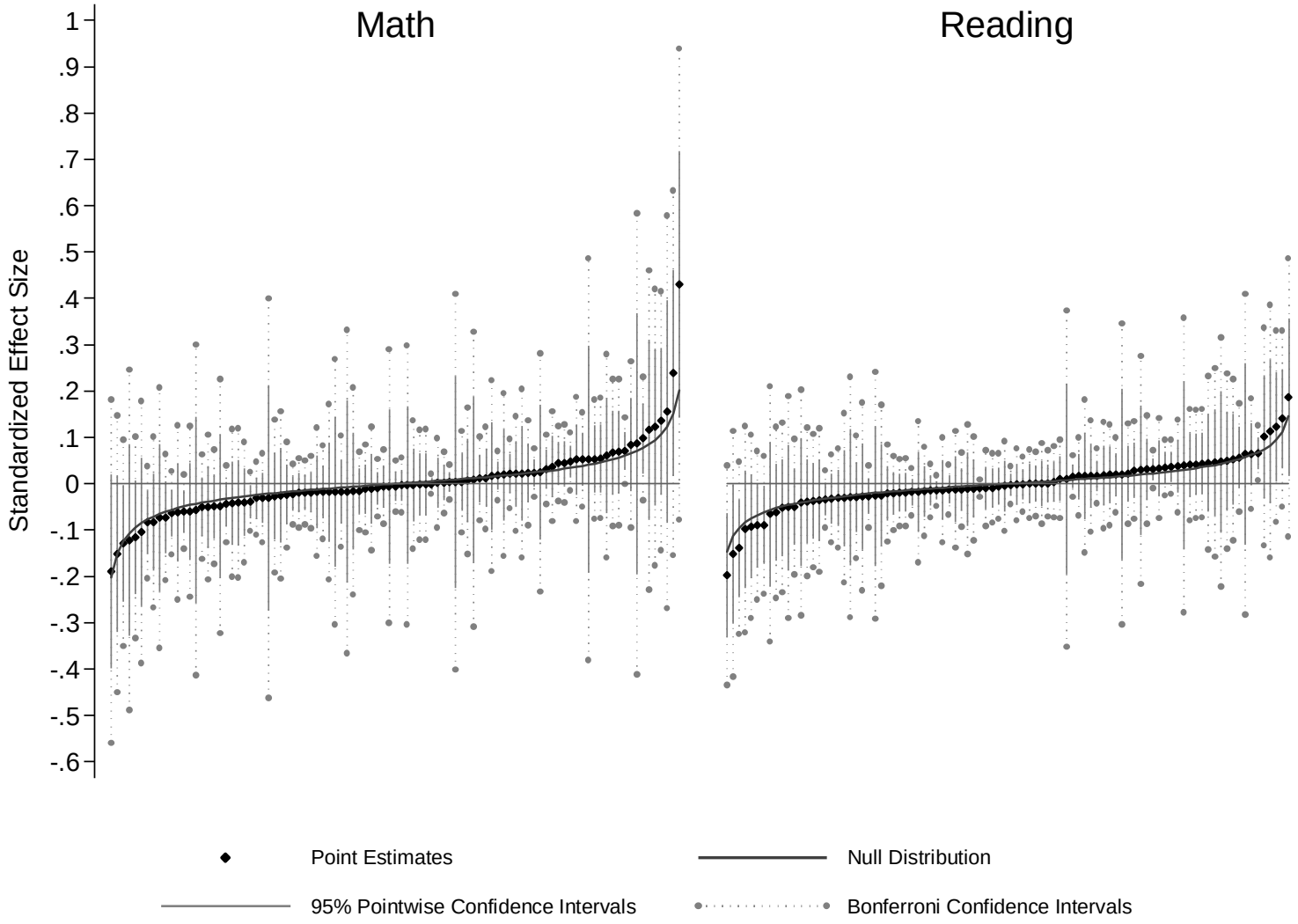

Figure 1. TPP contrasts from the all-grade models.

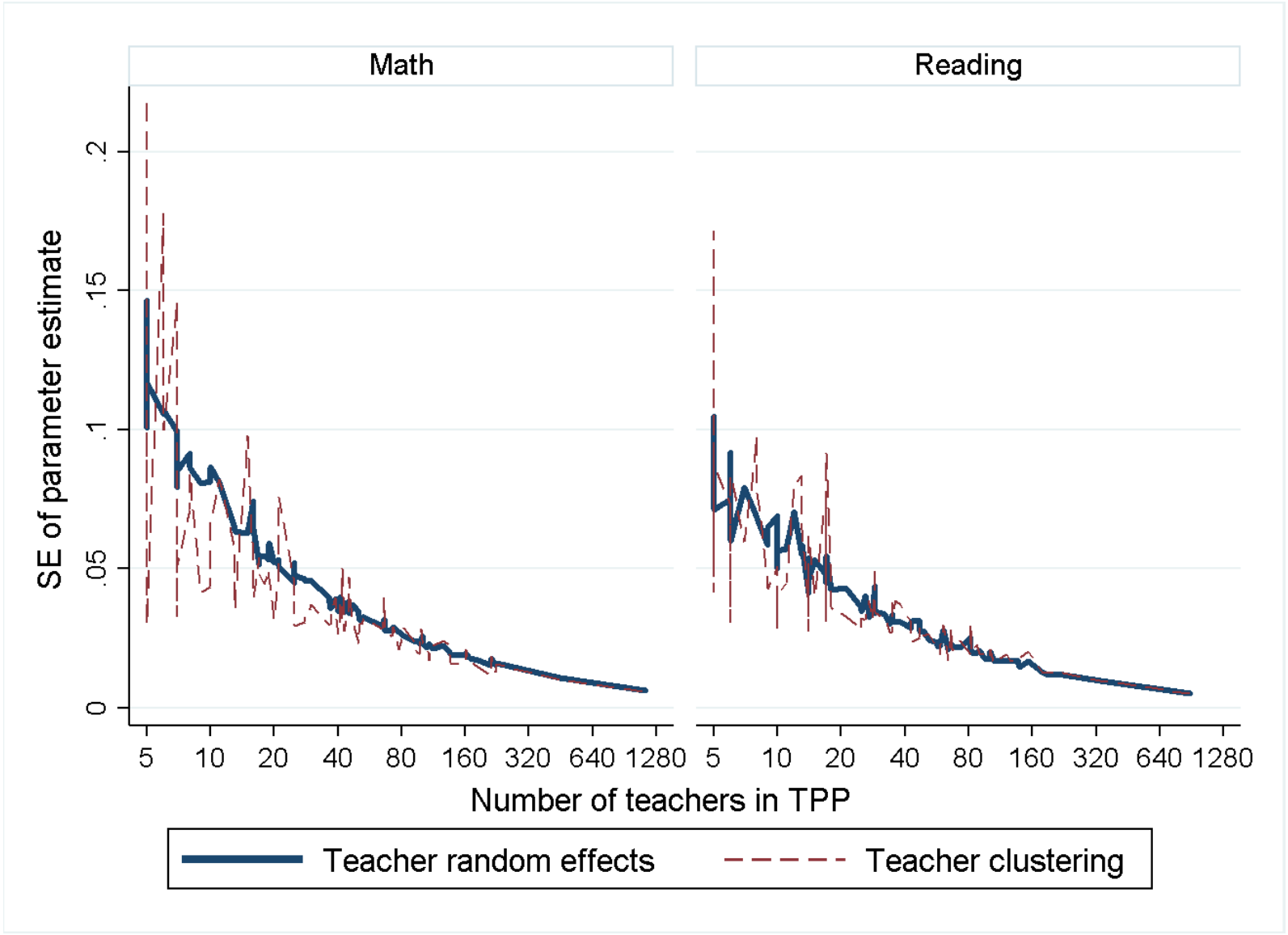

Figure 2. Standard errors (SE) of small and large TPPs under a model with teacher random effects vs. a model with teacher clustering.

Tables

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.

Table 2. TPPs in Texas, ranked by the number of 1st-3rd year math teachers in our data

TPP	Teachers Math	Reading
1. A+ Texas Teachers	1,067	823
2. iteachTEXAS	423	307
3. Region 04 Education Service Center	211	208
4. ACT-Houston	210	166
5. Texas A&M University	195	172
6. University of Texas - El Paso	195	141
7. Education Career Alternatives Program	171	132
8. Texas A&M University - Commerce	156	92
9. Texas State University-San Marcos	154	123
10. University of Texas - San Antonio	153	129
11. University of North Texas	146	114
12. Houston ISD	129	114
13. Sam Houston State University	127	95
14. Dallas ISD	124	80
15. Region 10 Education Service Center	110	77
16. Texas Tech University	108	96
17. Tarleton State University	103	65
18. Region 20 Education Service Center	100	79
19. West Texas A&M University	97	83
20. University of Texas - Austin	96	61
21. Region 01 Education Service Center	93	68
22. University of Houston	92	49
23. Texas Teaching Fellows (Dallas)	92	39
24. University of Texas - Pan American	88	72
25. Stephen F Austin State University	87	89
26. Texas Woman's University	76	60
27. ACT-San Antonio	71	39
28. ACT-Rio Grande Valley	66	54
29. Lone Star College	63	61
30. Web-Centric Alternative Cert Program	63	56
31. Texas A&M International University	63	33
32. University of Texas - Brownsville	60	56
33. Texas A&M University - Corpus Christi	51	35
34. University of Texas - Arlington	50	56
35. TeacherBuilder.com	46	29
36. Region 11 Education Service Center	46	53
37. Region 13 Education Service Center	45	16
38. Region 18 Education Service Center	44	33
39. Quality ACT	44	33
40. Texas Teaching Fellows (Austin)	44	29
41. A Career in Teaching-EPP (Corpus Christi)	43	27
42. Pasadena ISD	41	29
43. University of Houston-Victoria	41	29
44. Texas A&M University - Kingsville	41	41
45. Baylor University	38	24
46. Angelo State University	38	10
47. University of Houston-Clear Lake	38	43
48. Region 07 Education Service Center	37	13
49. Texas Alternative Certification Program	37	27
50. University of Houston-Downtown	36	44
51. LeTourneau University	35	31
52. Region 12 Education Service Center	32	27
53. Texas Teaching Fellows (San Antonio)	30	26
54. South Texas Transition to Teaching ACP	26	16
55. Lamar University	25	26
56. Texas A&M University - Texarkana	25	18
57. Lamar State College - Orange ACE Pgm	24	25
58. University of Texas - Tyler	24	55
59. Midwestern State University	23	14
60. Yes Preparatory Public Schools	22	13
61. University of Texas - Dallas	20	27
62. ACT-Houston at Dallas	20	24
63. Region 19 Education Service Center	19	9
64. University of Texas - Permian Basin	19	14
65. Wayland Baptist University	18	17
66. McLennan Community College	17	10
67. Lubbock Christian University	17	5
68. Alternative-South Texas Educator Program	15	15
69. University of the Incarnate Word	15	15
70. Sul Ross State University - Rio Grande	13	13
71. Collin County Community College	12	
72. Hardin-Simmons University	12	8
73. Region 03 Education Service Center	11	10
74. Texas Christian University	9	17
75. McMurry University	8	
76. Texas Alternative Center for Teachers	8	8
77. ACT-San Antonio at Central Texas	8	12
78. Southern Methodist University	8	8
79. Alamo Community College	8	14
80. St Mary's University	7	6
81. Howard Payne University	7	
82. Houston Baptist University	7	6
83. Region 14 Education Service Center	7	
84. Blinn College	6	7
85. Austin Community College	6	6
86. Texas Lutheran University	5	
87. University of Mary Hardin-Baylor	5	
88. Texas Wesleyan University	5	8
89. Region 05 Education Service Center	5	5
90. Our Lady of the Lake University	5	5
91. Austin College	5	
92. St Edward's University	4	4
93. Alternative Cert for Teachers NOW!	4	
94. Abilene Christian University	4	6
95. Educators of Excellence ACP	4	6
96. Dallas Baptist University		8
97. Southwestern Assemblies of God Univ		5
98. Alief ISD		5
99. A Career in Teaching-EPP (McAllen)		3
100. Concordia University		5

Table 3. Estimates, SEs, and significance of TPP effects

a. All TPPs

			Point estimates			Significantly different TPPs	
Subject	Model	TPPs	SD	Corr. with OLS	Mean of SEs	With Bonferroni correction	Without
Math	OLS	95	.071	1	.023	23	43
	OLS: teacher clustering		.071	1	.043	2	16
	OLS: school clustering		.071	1	.043	2	14
	OLS: district clustering		.071	1	.041	4	18
	OLS: TPP clustering		.071	1	.011	48	63
	RE: random teachers		.078	.89	.050	0	10
	RE: random teachers, schools		.075	.89	.050	0	7
	RE: random teachers, schools, districts		.075	.88	.050	0	5
Reading	OLS	92	.054	1	.027	5	28
	OLS: teacher clustering		.054	1	.039	1	10
	OLS: school clustering		.054	1	.039	1	10
	OLS: district clustering		.054	1	.037	3	14
	OLS: TPP clustering		.054	1	.011	37	56
	RE: random teachers		.056	.97	.041	1	11
	RE: random teachers, schools		.051	.95	.041	0	7
	RE: random teachers, schools, districts		.051	.95	.041	0	6

b. Large TPPs (≥40 teachers in subject)

			Point estimates			Significantly different TPPs	
Subject	Model	TPPs	SD	Corr. with OLS	Mean of SEs	With Bonferroni correction	Without
Math	OLS	48	.038	1	.010	17	27
	OLS: teacher clustering		.038	1	.025	1	9
	OLS: school clustering		.038	1	.026	0	8
	OLS: district clustering		.038	1	.026	1	9
	OLS: TPP clustering		.038	1	.007	25	35
	RE: random teachers		.036	.85	.026	1	7
	RE: random teachers, schools		.030	.78	.026	0	3
	RE: random teachers, schools, districts		.030	.75	.027	0	1
Reading	OLS	37	.022	1	.013	4	10
	OLS: teacher clustering		.022	1	.020	0	4
	OLS: school clustering		.022	1	.021	0	4
	OLS: district clustering		.022	1	.020	2	6
	OLS: TPP clustering		.022	1	.006	13	21
	RE: random teachers		.022	.84	.020	1	4
	RE: random teachers, schools		.021	.76	.020	1	3
	RE: random teachers, schools, districts		.020	.73	.021	0	2

Table 4. Comparing fit of all-grade TPP models

Subject	Model	District SD (SE)	School SD (SE)	Teacher SD (SE)	Residual SD (SE)	LR_{RE}
Math	OLS				.571*** (.001)	
	RE with random teachers			.194*** (.002)	.544*** (.001)	17,475***
	RE with random teachers, schools		.118*** (.005)	.157*** (.003)	.544*** (.001)	199***
	RE with random teachers, schools, districts	.040*** (.006)	.113*** (.005)	.157*** (.003)	.544*** (.001)	25***
Reading	OLS				.662*** (.001)	
	RE with random teachers			.119*** (.002)	.652*** (.001)	2,635***
	RE with random teachers, schools		.085*** (.004)	.086*** (.004)	.652*** (.001)	111***
	RE with random teachers, schools, districts	.023*** (.006)	.082*** (.005)	.086*** (.004)	.652*** (.001)	6**

*p<.05, **p<.01, ***p<.001. SEs in parentheses.
Note. OLS=ordinary least squares. RE=random effects. The LR_{RE} test compares each model to the model above it.

Table 5. Reliability and heterogeneity, estimated by comparing reading and math estimates

a. All TPPs (87 TPPs with both reading and math)

Estimates	Correlation (95% CI)			Heterogeneity SD $\hat{\tau}_{ICC}$ (95% CI)			*F*
OLS (with or without clustering)	.38	(.20	,.56)	.04	(.03	,.05)	2.14***
RE: random teachers	.43	(.26	,.60)	.04	(.03	,.05)	2.43***
RE: random teachers, schools	.42	(.24	,.59)	.04	(.03	,.05)	2.33***
RE: random teachers, schools, districts	.43	(.25	,.60)	.04	(.03	,.05)	2.39***

b. Large TPPs (36 TPPs with ≥40 teachers in both reading and math)

Estimates	Correlation (95% CI)			Heterogeneity SD $\hat{\tau}_{ICC}$ (95% CI)			*F*
OLS (with or without clustering)	.36	(.08	,.65)	.019	(.009	,.026)	1.98*
RE: random teachers	.31	(.01	,.61)	.017	(.003	,.024)	1.77*
RE: random teachers, schools	.31	(.01	,.61)	.015	(.002	,.021)	1.77*
RE: random teachers, schools, districts	.35	(.06	,.64)	.015	(.006	,.021)	1.93*

*p<.05, **p<.01, ***p<.001. CI=CI.

Table 6. Reliability and heterogeneity of all-grade estimates, estimated by comparing single-grade estimates

a. All TPPs

Subject	Estimates	$\hat{\rho}_{ICC}$ (95% CI)		$\hat{\tau}_{ICC}$ (95% CI)		F_{ICC}
Math	OLS (with or without clustering)	.40	(.15,.57)	.03	(.02,.04)	1.68***
	RE: random teachers	.42	(.18,.58)	.03	(.02,.04)	1.73***
	RE: random teachers, schools	.38	(.12,.55)	.03	(.01,.04)	1.62***
	RE: random teachers, schools, districts	.38	(.12,.55)	.03	(.01,.04)	1.61***
Reading	OLS (with or without clustering)	.26	(.00,.47)	.02	(.00,.03)	1.34*
	RE: random teachers	.36	(.08,.55)	.02	(.01,.03)	1.57**
	RE: random teachers, schools	.31	(.00,.51)	.02	(.00,.03)	1.45*
	RE: random teachers, schools, districts	.31	(.00,.51)	.02	(.00,.03)	1.44*

b. Large TPPs (≥40 teachers in subject)

Subject	Estimates	$\hat{\rho}_{ICC}$ (95% CI)		$\hat{\tau}_{ICC}$ (95% CI)		F_{ICC}
Math	OLS (with or without clustering)	.37	(0,.58)	.02	(0, .03)	1.59**
	RE: random teachers	.37	(0,.58)	.02	(0, .03)	1.60**
	RE: random teachers, schools	.19	(0,.46)	.01	(0, .02)	1.24
	RE: random teachers, schools, districts	.14	(0,.42)	.01	(0, .02)	1.16
Reading	OLS (with or without clustering)	.10	(0,.43)	.01	(0, .02)	1.11
	RE: random teachers	.11	(0,.44)	.01	(0, .02)	1.13
	RE: random teachers, schools	.10	(0,.43)	.01	(0, .02)	1.11
	RE: random teachers, schools, districts	.13	(0,.45)	.01	(0, .02)	1.15

Table 7. Tests of homogeneity

		All TPPs				Large TPPs	
Subject	Estimates	*df*	LR_{TPP}	*Q*	*W*	*df*	*Q*
Math	OLS	94	820***	819***	821***	47	567***
	OLS: teacher clustering			231***	232***		93***
	OLS: school clustering			250***	261***		87***
	OLS: district clustering			289***	518***		116***
	OLS: TPP clustering			3,265***	1×10^{8}***		1,297***
	RE: random teachers		148***	151***	150***		87***
	RE: random teachers, schools		122*	123*	124*		60
	RE: random teachers, schools, districts		117	117	118*		54
Reading	OLS	91	344***	344***	344***	36	169***
	OLS: teacher clustering			151***	153***		56*
	OLS: school clustering			154***	163***		54*
	OLS: district clustering			206***	502***		78***
	OLS: TPP clustering			2,586***	8×10^{7}***		847***
	RE: random teachers		134**	136***	136**		56*
	RE: random teachers, schools		119*	120*	120*		51
	RE: random teachers, schools, districts		111	112	112		45

*p<.05, **p<.01, ***p<.001. The LR_{TPP} is not reported for clustered models, because the calculation of the likelihood ignores clustering.

Table 8. Estimates of heterogeneity and reliability

a. All TPPs

		Reliability					Heterogeneity SD			
Subject	Estimates	$\hat{\rho}_H$	$\hat{\rho}_Q$	(95% CI)	$\hat{\rho}_{DL}$	$\hat{\rho}_{EB}$	$\hat{\tau}_H$	$\hat{\tau}_Q$	$\hat{\tau}_{DL}$	$\hat{\tau}_{EB}$
Math	OLS	.82	.89	(.87,.90)	.17	.19	.06	.07	.03	.03
	OLS: teacher clustering	.41	.59	(.49,.68)	.18	.09	.05	.05	.03	.02
	OLS: school clustering	.45	.62	(.53,.70)	.21	.12	.05	.06	.03	.02
	OLS: district clustering	.52	.67	(.60,.74)	.24	.15	.05	.06	.03	.03
	OLS: TPP clustering	.97	.97	(.97,.97)	.29	.72	.07	.07	.04	.06
	RE: random teachers	.40	.38	(.20,.52)	.08	.02	.05	.05	.02	.01
	RE: random teachers, schools	.38	.24	(.01,.41)	.04	.01	.05	.04	.02	.01
	RE: random teachers, schools, districts	.37	.20	(.00,.39)	.03	.00	.05	.03	.01	.00
Reading	OLS	.64	.73	(.67,.78)	.21	.13	.04	.05	.02	.02
	OLS: teacher clustering	.24	.39	(.22,.53)	.13	.04	.03	.03	.02	.01
	OLS: school clustering	.24	.41	(.24,.54)	.13	.04	.03	.03	.02	.01
	OLS: district clustering	.31	.56	(.44,.65)	.20	.09	.03	.04	.02	.02
	OLS: TPP clustering	.95	.96	(.96,.97)	.38	.74	.05	.05	.03	.05
	RE: random teachers	.31	.33	(.13,.48)	.09	.02	.03	.03	.02	.01
	RE: random teachers, schools	.18	.24	(.01,.42)	.07	.01	.02	.03	.01	.01
	RE: random teachers, schools, districts	.17	.19	(.00,.38)	.05	.01	.02	.02	.01	.00

b. Large TPPs (≥40 teachers in subject)

		Reliability					Heterogeneity SD			
Subject	Estimates	$\hat{\rho}_H$	$\hat{\rho}_Q$	(95% CI)	$\hat{\rho}_{DL}$	$\hat{\rho}_{EB}$	$\hat{\tau}_H$	$\hat{\tau}_Q$	$\hat{\tau}_{DL}$	$\hat{\tau}_{EB}$
Math	OLS	.92	.92	(.90, .93)	.49	.66	.04	.04	.03	.03
	OLS: teacher clustering	.50	.49	(.29, .64)	.26	.11	.03	.03	.02	.01
	OLS: school clustering	.49	.46	(.24, .62)	.24	.09	.03	.03	.02	.01
	OLS: district clustering	.44	.59	(.44, .70)	.36	.18	.03	.03	.02	.02
	OLS: TPP clustering	.95	.96	(.96, .97)	.49	.80	.04	.04	.03	.03
	RE: random teachers	.43	.46	(.24, .62)	.29	.12	.02	.02	.02	.01
	RE: random teachers, schools	.18	.22	(.00, .46)	.13	.02	.01	.01	.01	.00
	RE: random teachers, schools, districts	.09	.16	(.00, .42)	.09	.01	.01	.01	.01	.00
Reading	OLS	.63	.78	(.70, .84)	.78	.54	.02	.02	.02	.02
	OLS: teacher clustering	.04	.35	(.03, .57)	.32	.10	.00	.01	.01	.01
	OLS: school clustering	.01	.34	(.00, .56)	.30	.09	.00	.01	.01	.01
	OLS: district clustering	.07	.53	(.32, .68)	.61	.28	.01	.02	.02	.01
	OLS: TPP clustering	.91	.96	(.95, .96)	.92	.86	.02	.02	.02	.02
	RE: random teachers	.05	.36	(.04, .57)	.32	.10	.01	.01	.01	.01
	RE: random teachers, schools	.00	.29	(.00, .53)	.24	.06	.00	.01	.01	.01
	RE: random teachers, schools, districts	.00	.21	(.00, .48)	.16	.03	.00	.01	.01	.00