ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 20-330

# How Much Does Teacher Quality Vary Across Teacher Preparation Programs? Reanalyses from Six States

Paul T. von Hippel
University of Texas at Austin

Laura Bellows
Duke University

At least sixteen US states have taken steps toward holding teacher preparation programs (TPPs) accountable for teacher value-added to student test scores. Yet it is unclear whether teacher quality differences between TPPs are large enough to make an accountability system worthwhile. Several statistical practices can make differences between TPPs appear larger and more significant than they are. We reanalyze TPP evaluations from 6 states—New York, Louisiana, Missouri, Washington, Texas, and Florida—using appropriate methods implemented by our new caterpillar command for Stata. Our results show that teacher quality differences between most TPPs are negligible—.01-.03 standard deviations in student test scores—even in states where larger differences were reported previously. While ranking all a state's TPPs may not be possible or desirable, in some states and subjects we can find a single TPP whose teachers stand out as significantly above or below average. Such exceptional TPPs may reward further study.

VERSION: December 2020

Suggested citation: von Hippel, Paul T., and Laura Bellows. (2020). How Much Does Teacher Quality Vary Across Teacher Preparation Programs? Reanalyses from Six States. (EdWorkingPaper: 20-330). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/9mkt-pt42

A revision of this article was published in *Economics of Education Review*, *64*, 298-312.

# How Much Does Teacher Quality Vary Across Teacher Preparation Programs? Reanalyses from Six States

Paul T. von Hippel
University of Texas at Austin

Laura Bellows
Duke University

Draft of January 16, 2018

## Abstract

At least sixteen US states have taken steps toward holding teacher preparation programs (TPPs) accountable for teacher value-added to student test scores. Yet it is unclear whether teacher quality differences between TPPs are large enough to make an accountability system worthwhile. Several statistical practices can make differences between TPPs appear larger and more significant than they are. We reanalyze TPP evaluations from 6 states—New York, Louisiana, Missouri, Washington, Texas, and Florida—using appropriate methods implemented by our new *caterpillar* command for Stata. Our results show that teacher quality differences between most TPPs are negligible—.01-.03 standard deviations in student test scores—even in states where larger differences were reported previously. While ranking all a state's TPPs is not useful, in some states and subjects we can find a single TPP whose teachers are significantly above or below average. Such exceptional TPPs may reward further study.

*Correspondence*: Paul T. von Hippel, LBJ School of Public Affairs, University of Texas, 2315 Red River, Box Y, Austin, TX 78712, (512) 537-8112, paulvonhippel@utexas.edu

*Acknowledgment*: This research was supported by a grant from the Policy Research Institute at the University of Texas, Austin.

# 1 Introduction

Teacher preparation programs (TPPs) select, train, and certify public school teachers. While all public school systems require teacher preparation, TPPs differ substantially both in selectivity and in their approach to teacher training. Some TPPs accept as few 10 percent of applicants, while others take nearly all comers. Some TPPs are "traditional" 2- or 4-year degree programs, while others offer "alternative routes" which may require as little as 6 weeks' training before teachers begin their jobs. The lack of consistent and validated TPP standards has led to concerns about TPP quality. Some reformers have suggested that many TPPs are inadequate (Levine, 2006), and others have argued that TPPs are unnecessary, and that the teaching profession would improve if it opened to individuals who have not been trained by a TPP (Walsh, 2001).

In response to quality concerns, at least sixteen states have taken steps toward holding teacher preparation programs (TPP) accountable for teacher quality. The stated purpose of TPP accountability is to identify and "close failing [TPPs], strengthen promising programs, and expand excellent programs" (Levine, 2006; cf. US Department of Education, 2011). In addition, TPP quality ratings offer "consumer information" to "prospective teachers and employers (districts and schools)" as well as feedback to the "programs [TPPs] themselves" (Texas State Legislature, 2009).

Whereas traditional TPP accreditation emphasizes curriculum and faculty credentials, the new TPP accountability "focus[es] on student achievement as the primary measure of success" (Levine, 2006). Student achievement is estimated by test scores; teacher quality is estimated by value-added to test scores; and TPPs are held accountable for the average value-added by their teachers. While state TPP ratings may include several measures, teacher value-added typically receives substantial weight. Starting in 2010, the federal government provided grants to help some states rate TPPs in this manner, and 3 days before the 2016 election the US Department of Education issued a rule requiring that all states do so (Department of Education, 2016). But 4 months after the election, Congress repealed the new rule (115th Congress, 2017).

Is this form of TPP accountability constructive, or worthy of repeal? The motivation behind TPP accountability seems very plausible at first. Teachers vary in value-added—one standard deviation (SD) in teacher value-added equals about 0.1 SD in student test scores—and TPPs vary both in selectivity and in their approach to teacher training. It stands to reason that some TPPs would turn out better teachers than others, either because the better TPPs select trainees who have exceptional potential, or because the better TPPs provide exceptional training.

It does not necessarily follow, though, that the differences between teachers from different TPPs are large enough to warrant policy action. Indeed, in many professions, little of the variation in productivity lies between workers selected and trained by different institutions. Among PhD economists, only 10 percent of the variance in research productivity lies between graduates of different PhD programs (Conley & Önder, 2014).[1] Among college graduates with the same major, only 1 to 9 percent of the variance in log earnings lies between graduates of different colleges (Rumberger & Thomas, 1993). Among teachers, if a similar percentage of the variance in value-added lies between graduates of different TPPs, then a back-of-the-envelope calculation[2] suggests that the SD between TPPs would amount to just .01-.03 SDs in student test scores.

[1] We calculated this fraction of variance by running an ANOVA on data published by Conley and Önder (2014). Conley and Önder summarize their results in a different way.

[2] We get this figure by multiplying the SD of teacher value-added, which is about .1 SD in student test scores, by the square root of 1 to 10 percent, which is the percentage of variance in productivity that typically lies between workers trained by different institutions. Then .1 SD×($.01^{1/2}$ to $.10^{1/2}$)= .01 to .03 SD.

Differences of this size are not just small; they can be practically impossible to estimate with any certainty. One problem is estimation error; effects of .01-.03 SD are usually small compared to their standard errors (SEs), and may also be small compared to minor biases that result from the misspecification of value-added models.

Another problem is *multiple comparisons*. In Texas, for example, there are approximately 100 different TPPs, and if we test each of them using a .05 significance level, we would expect to conclude that approximately five differ significantly from the average—even if all are in fact identical. Even in a smaller state with 10 identical TPPs, ordinary hypothesis tests would run a 40 percent chance ($1–(1-.05)^{10}$) of erroneously concluding that at least one TPP differs significantly from the average. Although most TPP evaluations have neglected the issue of multiple comparisons, it is appropriate to correct significance levels and CIs for the number of TPPs being compared. After correction, few if any TPPs may differ significantly from the average (von Hippel, Bellows, Osborne, Lincove, & Mills, 2016).

In addition to these fundamental challenges, a number of choices made in analysis can exaggerate apparent differences between TPPs. The Methods section will discuss these choices in detail, but in brief they include underestimation of SEs, display of narrow confidence intervals (CIs) that extend only one SE in each direction, underappreciation of how noise affects the distribution of TPP estimates, and confounding of between-TPP variance with variance in a comparison group of experienced teachers.

## *1.1 Empirical review*

Results reported from past TPP evaluations are confusingly mixed. In some states, results have been consistent with our discussion, suggesting that there are only trivial differences between teachers from different TPPs, and that it is rarely possible to tell which TPPs are better or worse (Koedel, Parsons, Podgursky, & Ehlert, 2015; von Hippel et al., 2016). Yet in other states, evaluators have concluded that the differences between TPPs are more substantial, and that it is practical to single out TPPs whose teachers are better or worse than average (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2009; Gansle, Noell, & Burns, 2012).

While it is possible that the true differences between TPPs are larger in some states than in others, it is also possible that these differences are more apparent than real. The results of TPP evaluations in different states may vary not for substantive reasons, but because of the methodological choices made by different states' evaluators. It is also possible that the messages of different evaluations differ not because of the statistical results *per se*, but because of the way that they have been interpreted. Faced with the same set of results, some evaluators may believe they see intriguing differences between TPPs, while others may conclude that the true differences are small, and that any apparent differences consist mostly of estimation error, or noise.

Until now it has been difficult to know to what extent the differences between TPP evaluations result from differences in substance, methods, or interpretation. While recent articles have raised concerns about the methods used to evaluate TPPs in some states (Koedel & Parsons, 2014; Koedel et al., 2015; von Hippel et al., 2016), it has been difficult to evaluate these concerns empirically, because TPP evaluations typically use restricted state data which is not available for reanalysis.

In this article, we reanalyze the results of TPP evaluations from 6 states: Louisiana, Missouri, Washington, Texas, Florida, and New York (City). We can do this because our statistical methods do not require access to the original data. Instead, our methods, which are similar to those used in meta-analysis, only require point estimates and SE estimates—statistics that are commonly available in published tables

and graphs.[3] Our methods are implemented in our new *caterpillar* command, which can be installed in Stata by typing *ssc install caterpillar, all*. Installation of *caterpillar* will also download data and code that replicates nearly all of the results in this article.

Our reanalyses clear up most of the apparent discrepancies. In every state, our results suggest that teacher quality differences between most TPPs are negligible—even in Louisiana and New York City, where larger differences were reported originally. On review, it appears that differences between TPPs are rarely detectible, and that if they could be detected they would usually be too small to support effective policy decisions. That said, in some states and subjects, we can occasionally identify a single TPP that is significantly different from the average—and in one state the size of the difference is not trivial in size.

A limitation of the reviewed studies is that they rely on test scores. Test scores proxy for students' academic skills, and there is evidence that teachers who raise test scores also improve later outcomes such as high school graduation, college graduation, earnings, and wealth (Chetty, Friedman, & Rockoff, 2014; Koedel, 2008). Nevertheless, when stakes are attached, some teachers may find ways to raise average scores without commensurate improvement in skills or later outcomes (Koretz, 2002, 2009; Quezada-Hofflinger & von Hippel, 2017). We should be careful to ensure that accountability systems do not encourage TPPs and teachers to game the test.

One recent study evaluated TPPs using principals' ratings of teachers, and found more heterogeneity than we find using test scores (Ronfeldt & Campbell, 2016). While this finding is intriguing, it is unclear whether principal ratings predict future student success, as test scores do. In addition, principal ratings are biased in favor of teachers who teach advantaged students, and biased in favor of teachers whom a principal has evaluated positively in the past (Steinberg & Garrett, 2016; Whitehurst, Chingos, & Lindquist, 2014). While the bias toward advantaged students can be addressed with student covariates (Ronfeldt & Campbell, 2016),[4] the bias toward favored teachers is harder to address, and raises the concern that a halo effect may inflate the evaluations of teachers hired from a principal's favorite TPPs.

# 2 Methods

A TPP evaluation begins with a value-added model which estimates the average effect of each TPP's teachers on student test scores. Next, TPP estimates from this model can be *post-processed* to determine how much of the variation across TPP estimates is due to *heterogeneity* (true differences) among TPPs, rather than estimation error. In addition, hypothesis tests can try to single out which individual TPPs differ significantly from the average.

## *2.1 Value-added model*

TPP evaluations typically fit a *lagged-score value-added model* which, for student *i* taught by teacher *j* in year *t*, regresses that year's test score $Y_{ijt}$ in some subject (e.g., math) on some function *f* of the prior year's score $Y_{i(t-1)}$, or possibly a vector of scores $\boldsymbol{Y}_{i(t-1)}$ in different subjects (e.g., math and reading). The model also includes a vector of covariates $\boldsymbol{X_i}$ measured at multiple levels (e.g., student, teachers, classroom, school, district), and a vector of dummies $\boldsymbol{TPP}_j$ indicating which TPPs each teacher *j* attended:

[3] When estimates are not available in published form, we obtained them from the evaluators.

[4] In their evaluation of TPPs with principal ratings, Ronfeldt and Campbell (2016) controlled for student body characteristics at the school level, but not at the classroom level.

$$Y_{ijt} = \boldsymbol{\alpha}\, \boldsymbol{TPP}_j + f(\boldsymbol{Y}_{i(t-1)}) + \boldsymbol{\beta_2 X_i} + e_{ijt} \quad (1).$$

The TPP coefficients are in the vector $\boldsymbol{\alpha}$.

Within this basic framework, different covariates can be used, different functional forms can be assumed, and different decisions can be made about fixed effects, random effects, and comparison groups. Some of these decisions have implications for the TPP estimates and their SEs, which we review next.

### 2.1.1 Teacher clustering or teacher REs

The residual $e_{ijt}$ should be structured to reflect the correlations among different students taught by the same teacher (Koedel et al., 2015). One option is to cluster SEs at the teacher level. Another option is to add teacher random effects (REs), which split $e_{ijt}$ into two components $e_{ijt} = r_j + u_{ijt}$, where $r_j$ is a teacher RE, $u_{ijt}$ is a random residual, and $r_j$ and $u_{ijt}$ are uncorrelated with each other and with the regressors.

Note that the clusters or REs should be at the teacher level rather than the classroom level. One teacher can teach several classrooms, and if the clusters or REs are at the classroom level, then different classrooms taught by the same teacher will be treated as independent, and the SEs of the TPP estimates will be too small (Koedel et al., 2015).

Teacher clustering and teacher REs produce similar estimates in large TPPs, but in small TPPs teacher REs are preferable, since teacher clustering produces SE estimates that are volatile and biased downward (too small). There is no hard line between small and large TPPs, but for TPPs with at least 40 teachers the bias and volatility in SEs appear negligible (von Hippel et al., 2016). This finding is consistent with the general result that ordinary clustered SEs for treatment effects are volatile and biased downward if a treatment group has fewer than 40 clusters (MacKinnon & Webb, 2017). Accordingly, in each state we report separate results for all TPPs and for large TPPs with at least 40 teachers in the data.[5]

An alternative way to account for the nesting of students within teachers is to fit the TPP model in two stages (Goldhaber, Liddle, & Theobald, 2013). Stage 1 estimates *teacher* value-added by fitting regression (1) with a dummy for each teacher instead of each TPP. Stage 2 regresses teacher value-added on TPP dummies and covariates. A two-stage model can produce consistent estimates, but it is less efficient than fitting a single model with teacher clusters or teacher REs (Raudenbush & Bryk, 2001).

### 2.1.2 School REs and school FEs

Above the teacher level one can add school REs, which slightly improve model fit and slightly increase SEs (von Hippel et al., 2016). One can also use school fixed effects (FEs) (Boyd et al., 2009), but this can introduce problems. School FEs are meant to reduce bias by accounting for between-school differences that are not captured by covariates. But school FEs reduce the estimation sample to schools employing recent graduates from multiple TPPs. The reduced sample will produce less efficient TPP estimates, even no estimates at all for some small TPPs. In addition, reducing the sample may introduce bias since the schools in the reduced sample are nonrepresentative—larger, higher turnover schools with multiple vacancies (Mihaly, McCaffrey, Sass, & Lockwood, 2013; von Hippel et al., 2016).

[5] In an evaluation with several years of data, a "large" TPPs might turn out just 10 or so teachers per year, but cumulatively turn out 40 or more over the years of the study. Note that the data are limited to teachers in the grades and subjects where tests are given and value-added scores are calculated.

### 2.1.3 Experienced comparison group

While most TPP evaluations limit the sample to recent TPP graduates, it is possible to include a comparison group of more experienced teachers whose TPPs are not known (Gansle et al., 2012; Mihaly et al., 2013). The idea is attractive since it increases the size of the estimation sample, so that the coefficients of covariates can be estimated more precisely. And if school FEs are used, including experienced teachers means that it is no longer necessary to discard much of the sample, since any recent TPP graduate will be included provided there is an experienced teacher in the same school.

But bias can result if experienced teachers are included in a model that also includes school FEs. Without school FEs, recent TPP graduates are compared to experienced teachers as a group, but with school FEs, recent TPP graduates are compared to the experienced teachers *who work in the same schools*. Since the quality of experienced teachers varies across schools, the between-school variation among experienced teachers will be confounded with between-school variation among recent TPP graduates. (Similar though smaller biases can result from the inclusion of other comparison groups such as teachers who were certified out of state.)

The bias resulting from an experienced comparison group can be either positive or negative. If a TPP sends its graduates to schools whose experienced teachers are relatively effective, then the TPP graduates' effectiveness will be underestimated relative to other TPPs. But if a TPP sends its graduates to schools whose experienced teachers are relatively *in*effective, then the TPP graduates' effects will be *over*estimated relative to other TPPs.[6] Similar though smaller biases may arise in models with school REs, since estimates from school RE models are a weighted average of OLS estimates and estimates from school FE models (Greene, 2011; von Hippel et al., 2016).

## *2.2 Post-processing of point estimates and SEs*

Estimating TPP coefficients is only the first step. Next we must estimate how large, reliable, and significant the true differences between TPPs are. We do this using methods that are similar to those used in meta-analysis. The advantage of these methods is that they only need point estimates and SEs, which are typically all we have when reanalyzing the results of a published TPP evaluation. Our methods are implemented in our *caterpillar* command, which can be installed in a Stata session by typing *ssc install caterpillar*.

### 2.2.1 SE-based estimates of heterogeneity and reliability

From the value-added model in (1), or a variant, we get a vector of coefficient estimates $\hat{\boldsymbol{\alpha}} = [\hat{\alpha}_1 \ldots \hat{\alpha}_P]$ for each of the $P$ TPPs. Each estimate $\hat{\alpha}_p$ has a standard error estimate $s_p$. We center the

[6] To put the argument more formally, let the coefficient $\alpha_p$ represent the average value-added by recent graduates from the $p^{\text{th}}$ TPP. With school FEs, what we are actually estimating is not $\alpha_p$ but $\alpha_p - \gamma_p$, where $\gamma_p$ is the average value-added by experienced teachers who work in the same schools as teachers from the $p^{\text{th}}$ TPP. Across TPPs, the variance of $\alpha_p - \gamma_p$ is not the same as the variance of $\alpha_p$. To the contrary, $Var(\alpha_p - \gamma_p) = Var(\alpha_p) + Var(\gamma_p) - 2\,Cov(\alpha_p, \gamma_p)$, so if $Cov(\alpha_p, \gamma_p)$ is negligible (as we suspect it is), then the variance of $\alpha_p - \gamma_p$ will exceed the variance of $\alpha_p$. If $Cov(\alpha_p, \gamma_p)$ is negative—implying that stronger TPPs send their graduates to schools with weaker experienced teachers—then the variance of $\alpha_p - \gamma_p$ will increase further. But if $Cov(\alpha_p, \gamma_p)$ is positive—implying that stronger TPPs send their graduates to schools with stronger experienced teachers—then the variance of $\alpha_p - \gamma_p$ will be somewhat reduced.

estimates around their precision weighted mean $\bar{\alpha} = \sum(s_p^{-2}\hat{\alpha}_p) / \sum s_p^{-2}$ to estimate TPP *contrasts* $\Delta\hat{\alpha}_p = \hat{\alpha}_p - \bar{\alpha}$.

Some of the variance among the TPP contrasts is due to *heterogeneity*—i.e., differences among the true TPP coefficients $\alpha_p$. But some of the variance is due to random estimation error, which is reflected in the standard errors $s_p$. By the law of total variance we have

$$V(\Delta\hat{\alpha}_p) = V(\hat{\alpha}_p) = \tau^2 + E(s_p^2) \quad (2),$$

where $\tau^2 = V(\alpha_p)$ is the heterogeneity variance. If there is no heterogeneity, then the TPPs are *homogeneous*. The null hypothesis of homogeneity is $H_0: \tau^2 = 0$.

A test of the null hypothesis is Cochran's (1954) *Q* statistic (cf. Koedel et al., 2015):

$$Q = \sum \Delta\hat{\alpha}_p^2 \, s_p^{-2} \quad (3)$$

Under the null hypothesis, *Q* follows a $\chi^2_{P-1}$ distribution if the SE estimates $s_p$ are accurate—i.e., if the value-added model is correctly specified and the number of teachers per TPP is reasonably large.

*Q* is easy to calculate when all we have are point estimates and SEs, yet *Q* is also powerful and robust. Despite its simplicity, *Q* is almost identical to the likelihood ratio statistic *LR*, which is the uniformly most powerful test, calculated by comparing the likelihoods of model (1) with and without the TPP indicators (von Hippel et al., 2016). It is true that *Q* ignores the correlations among the estimates $\Delta\hat{\alpha}_p$—i.e., the off-diagonal terms of the covariance matrix whose diagonal terms are $s_p^2$—but this may make *Q* more robust. In one TPP evaluation, an alternative to *Q* that included off-diagonal elements typically returned results very similar to *Q*, but occasionally "blew up" (took implausibly large values) when the off-diagonal elements were estimated with too few clusters (von Hippel et al., 2016, n. 6).[7]

From (2) we can derive a consistent and unbiased estimate of the heterogeneity variance

$$\hat{\tau}^2 = \hat{V}(\hat{\alpha}_p) - \frac{1}{P}\sum s_p^2 \quad (4)$$

where $\hat{V}(\hat{\alpha}_p)$ is the sample variance of the TPP estimates (Cochran, 1954). Also

$$\hat{\rho} = 1 - \frac{P-1}{Q} \quad (5)$$

consistently estimates the *reliability* $\rho = \tau^2/V(\hat{\alpha}_p)$, which is the fraction of variance in the TPP estimates that is due to heterogeneity rather than estimation error (Higgins & Thompson, 2002; Koedel et al., 2015). Both $\hat{\tau}^2$ and $\hat{\rho}$ can take negative values, which are rounded up to zero. Rounding upward induces a slight positive bias when there are few TPPs and little or no true heterogeneity (von Hippel, 2015).

*Q*, $\hat{\tau}^2$, and $\hat{\rho}$ are convenient statistics because they can be calculated from point estimates and SEs. Alternative statistics typically require information that is not available in reanalysis, and are no better—or in some cases worse—than the statistics we use here. For example, some evaluations have estimated $\tau^2$ as the variance of empirical Bayes (EB) TPP estimates (e.g., Boyd et al., 2009; Goldhaber et al., 2013; Koedel et al., 2015), but this estimate has a large negative bias. To see the bias simply, remember that if all TPP estimates are equally reliable[8] the EB estimates are $\Delta\hat{\alpha}_p\tilde{\rho}$, where $\tilde{\rho}$ is a consistent estimate of reliability. Then the variance of the EB estimate is $\hat{\tau}^2_{EB} = \tilde{\rho}^2 Var(\Delta\hat{\alpha}_p)$, whereas the true heterogeneity is

[7] We know that clustered SEs are biased and volatile when there are too few clusters, and we speculate that the bias and volatility are more severe for the off-diagonal terms of the covariance matrix.

[8] If each estimate has a different reliability, the EB variance formula is more complicated but still biased.

$\tau^2 = \rho Var(\Delta\hat{\alpha}_p)$. So the variance $\hat{\tau}^2_{EB}$ of the EB estimates underestimates $\tau^2$ by approximately a factor of $\rho$ (von Hippel et al., 2016, p. 37). None of the analyses in this paper use EB estimates.

### 2.2.2 Graphical techniques

TPP estimates are often presented visually, with "caterpillar plots" that graph TPP estimates or contrasts, sorted from smallest to largest. These caterpillar plots typically have a sideways S shape (see Figure 1, for example), which has led to conventional wisdom suggesting that, while most TPPs are practically indistinguishable, a few very good or very bad ones stand out in the tails.

But to interpret a caterpillar plot properly, we must compare the distribution of observed TPP estimates with the *null distribution* that would be expected under the null hypothesis of homogeneity, when estimated TPP differences would consist of nothing but estimation error. The null distribution is an equally weighted mixture of $P$ normal mixture distributions, $N(0, s_p^2)$, $p = 1,\ldots, P$ (von Hippel et al., 2016, p. 36)—and as it happens, this mixture is also a sideways S; it looks like a cumulative normal distribution, but turned on its side and with heavier tails. So a sideways S distribution does not, by itself, tell us anything about the distribution of true TPP effects; a sideways S would be expected even if all TPPs were identical (von Hippel et al., 2016, p. 39).

Our *caterpillar* command for Stata produces an enhanced caterpillar plot which lays the null distribution over the observed distribution of point estimates. If the null distribution fits the observed distribution, then it would appear that the TPPs are nearly homogeneous. But if the observed estimates depart from the null distribution, then we have evidence for heterogeneity, and we may even be able to tell which individual TPPs are responsible for it.

For policy purposes—if we wish to "close failing [TPPs]…and expand excellent programs" (Levine, 2006)—it is not enough to know whether heterogeneity is present. We must also single out the specific TPPs are better or worse than average. To do this, it is common to plot the $P$ contrasts $\Delta\hat{\alpha}_p$ with pointwise CIs, and see which CIs do not cover zero; those TPPs are interpreted as differing significantly from the mean. The problem with this approach is that it ignores the problem of multiple comparisons (Hsu, 1996). In Texas, for example, there are approximately $P$=100 TPPs, and if all were identical we would expect that five 95 percent CIs would fail to cover zero. The problem of multiple comparisons is exacerbated if we graph 68 percent CIs that extend only one SE in each direction (Boyd, Grossman, Lankford, Loeb, & Wyckoff, 2009; Gansle, Noell, & Burns, 2012a). If even $P$=10 identical TPPs are compared using 68 percent CIs, there is a 98 percent chance ($1–.68^{10}$) of erroneously concluding that at least one TPP differs significantly from the average.

The Bonferroni correction adjusts for multiple comparisons by producing wider CIs each of which has a confidence level of ($100–5/P$) percent, so that when all $P$ CIs are considered together, the chance of erroneously concluding that at least one of them differs from average is no more than 5 percent. The Bonferroni correction is slightly conservative, and less conservative corrections are available, including one tailored for our problem of making multiple comparisons with the mean (Fritsch & Hsu, 1997). But the exact correction is complicated, and if the sample is not too small, the exact result is practically indistinguishable from the simple Bonferroni correction. For example, with $P$=10 TPPs and at least five teachers per TPP, a 95 percent Bonferroni CI is only 1 percent wider than the exact CI (Fritsch & Hsu, 1997, Table 13.1). Our results use the Bonferroni correction; using the exact correction would not visibly change the results.

We correct for the $P$ comparisons made within a single subject and state. This is most appropriate for middle and high school teachers who are certified to teach a single subject. In training teachers for middle and high school, TPPs may offer different curricula in different subjects, which can be held

accountable separately. In elementary school, where teachers often receive a general certification, it might be more appropriate to correct for all subjects together. Naturally, correcting for a larger family of comparisons would yield wider confidence intervals and fewer significant differences.[9]

A related issue is that, while TPP estimates may in general be unbiased, if we highlight only those estimates that pass a significance test, the highlighted estimates will typically be exaggerated (Gelman, 2017). While TPP estimates are sometimes shrunk according to their reliability (Section 2.2.1), this practice may not suffice to eliminate the bias that come from selecting on significance.

### 2.2.3 Correlation-based estimates of heterogeneity and reliability

All the approaches so far depend on SEs, and this can cause bias since SE estimates are often biased in TPP evaluations. Another approach, which does not require SE estimates, is to estimate heterogeneity and reliability by comparing different point estimates of the same TPP effects. For example, if we have TPP estimates for both 6th and 7th grade math teachers, their correlation estimates reliability $\rho$, and their covariance estimates the heterogeneity variance $\tau^2$. If there are more than two sets of estimates—e.g., for 6th, 7th, and 8th grade math teachers—the correlation generalizes to the intraclass correlation (i.e., the intra-TPP correlation), and the covariance generalizes to the intraclass covariance, better known as the between-TPP variance. These statistics can be estimated by fitting the ANOVA model:

$$\Delta\hat{\alpha}_{pj} = \Delta\alpha_p + u_{pj} \qquad (6),$$

where $\Delta\hat{\alpha}_{pj}$ is the $j^{th}$ contrast estimate for TPP $p$, $\Delta\alpha_p$ is the true contrast, and $u_{pj}$ is random estimation error. Here the between-TPP variance is $\tau^2 = V(\Delta\alpha_p)$ and the intraclass correlation is the reliability $\rho = \tau^2/V(\Delta\hat{\alpha}_{pj})$. In Stata, these statistics are most conveniently calculated using the *loneway* command.

Even if two sets of TPP estimates do not refer to the same effects, it can still be informative to check their correlation and covariance. For example, we do not necessarily expect TPPs to have exactly the same effect in reading and math, but if heterogeneity is present we would expect there to be some positive correlation between a TPP's reading and math estimates. This correlation may underestimate reliability if teachers from the same TPP have different value-added in reading than in math. But it may overestimate reliability if reading and math value-added are estimated from the same teachers, who do not provide independent evidence about their TPPs.

### 2.2.4 Model uncertainty

Evaluators sometimes report the correlation between TPP estimates from different models fit to the same scores (e.g., Goldhaber et al., 2013; Koedel et al., 2015). This does not estimate reliability in the sense described earlier, since the estimates are not independent. But it does highlight the sensitivity of TPP estimates to modeling decisions. In Washington state, for example, the correlation between TPP estimates from different models was over .9 if the models were similar, but could be .1 or lower if the models were very different (Goldhaber et al., 2013, Table 6).

Given the variety of TPP models that have been fit, and the fact that every model is to some degree misspecified, it would be reasonable to fit a variety of models and combine the results (with better-fitting models getting more weight) to produce TPP point estimates and SEs that reflect model uncertainty (Burnham & Anderson, 1998). To our knowledge, no one has done this in the value-added literature, but the resulting SEs would be larger than the SEs that are commonly reported.

---

[9] It makes no sense to correct for comparisons across different states. Accountability in Missouri, for example, should not depend on comparisons made in Texas.

# 3 Data and results for individual states

In peer-reviewed journals, TPP evaluations have been published for six different states: Texas, Washington state, Missouri, New York (City), Florida, and Louisiana. For each state, we obtained and analyzed TPP point estimates and SEs ($\hat{\alpha}_p$ and $s_p^2$). From these we calculated estimates and tests of heterogeneity and reliability. We also flagged the TPPs that were or were not significantly different from average, both before and after adjustment for multiple tests.

For some states, the authors provided us with estimates by email; for others, we copied estimates from published tables and figures. To facilitate comparisons, we standardized all estimates to a scale where one unit was equal to 1 SD in student test scores. We have made our TPP estimates available in a replication dataset, which you can download in Stata by typing *ssc install caterpillar, all*. The replication dataset includes every state but Missouri, whose data sharing agreement did not permit publication of individual TPP estimates.

## *3.1 Texas*

We were part of a team that evaluated TPPs in Texas (von Hippel et al., 2016). Texas TPPs are highly diverse; "although many…are traditional programs run out of colleges and universities, the…four largest TPPs are alternative TPPs, three of which are run for profit" (von Hippel et al., 2016). The evaluation estimated value-added for 95 TPPs in math and 92 in reading. Texas TPPs varied substantially in size, with the largest TPP contributing over 1,000 teachers, and the smallest contributing less than 5. One analysis included all TPPs, and another was limited to teachers from larger TPPs, defined as TPPs that contributed at least 40 teachers to the data. By this definition, there were 48 large TPPs in math and 37 in reading.

Value-added estimates were limited to a single school year, 2010-11, and produced separate value-added estimates for math in each grade from 4 to 10 and for reading in each grade from 4 to 9. The evaluation was limited to teachers with 0-2 years of prior experience—6,358 teachers linked to nearly 300,000 student math scores, and 4,965 teachers linked to over 200,000 student reading scores.

The evaluation fit the lagged-score value-added model in (1). The left side was a standardized test score in math or ELA, and the right side included lagged test scores in both subjects; student, classroom, school, and district covariates; grade dummies; and the teachers' years of experience. We present results from the model with REs at both the teacher and school level, which fit better than a model with REs at the teacher level alone. The Texas evaluators also fit models with teacher- or school-clustered SEs; these models gave similar estimates for large TPPs, but gave volatile and biased SE estimates for small TPPs (von Hippel et al., 2016, Figure 2).

Figure 1 summarizes the distribution of TPP estimates in Texas. The results suggest little heterogeneity. The null distribution fits the point estimates very well—so well that it is worth repeating how the null distribution was calculated. The null distribution was *not* calculated by fitting a curve to the point estimates; in fact, it did not use the point estimates at all. Instead, calculation of the null distribution used the SE estimates to calculate what the distribution of point estimates *would* look like under homogeneity. The fact that the null distribution fits the point estimates so well suggests that the TPPs are nearly homogeneous—i.e., that there is little true difference between the value-added by teachers from different TPPs.

The statistics in Figure 1 confirm the impression that little heterogeneity is present. The *Q* test rejects homogeneity in both reading and math, but the *p* values are not very small ($.02<p<.03$) considering the large sample size and high power. The estimates are just 23-25 percent reliable, and no TPP differs significantly from the average after adjustment for multiple tests. Among large TPPs, the *Q* tests provide

only borderline evidence against homogeneity in both reading and math ($.05<p<.10$); the heterogeneity estimate is just .00-.02 student-level SDs, and only one TPP differs significantly from the average in reading—none in math.

Instead of relying on the SE estimates, an alternative is to estimate the correlation and the square root of the covariance between reading and math point estimates. Among all TPPs, these estimates give a reliability of 42 percent and a heterogeneity SD of .04 (Table 1b). Among large TPPs, they give a reliability of 30 percent and a heterogeneity SD of .01 (Table 2b). These estimates are highly significant ($p<.001$), but do little to help us identify individual TPPs that are significantly better or worse than average.

## *3.2 Washington state*

An evaluation in the state of Washington produced estimates for the state's traditional university-based TPPs (Goldhaber et al., 2013). The evaluation drew data from 2 school years and included data on 6,827 $4^{th}$-$6^{th}$ grade teachers from 20 Washington TPPs, plus a comparison group of 1,891 teachers who were trained in other states. The Washington evaluators fit the two-stage lagged-score value-added model described in Section 2.1.1, and clustered the SEs, appropriately, at the teacher level (Goldhaber et al., 2013, n. 33). We copied TPP estimates directly from Table 4 in the Washington evaluation. Several models were estimated; we chose the one that was most similar to models fit in other states (Goldhaber et al., 2013, Table 4, model 1).

Figure 2 summarizes the TPP contrasts for Washington state. There is no evidence of heterogeneity in math, where the null distribution fits the point estimates almost perfectly and the *Q* test is nonsignificant ($p>.4$ among the large TPPs and among all TPPs). There is evidence of heterogeneity in reading, where the *Q* test is significant ($p\leq.001$ among the large TPPs and among all TPPs). But even in reading there is only one TPP that differs significantly from the average after adjustment for multiple tests. That TPP is at the far left of the caterpillar plots for reading; its point estimate suggests that its teachers' value-added to student test scores is just .05 SD worse than average, and this estimate is likely exaggerated by the fact that we selected it on grounds of statistical significance (Gelman, 2017).

Instead of relying on SE estimates, an alternative is to estimate the correlation and the square root of the covariance between reading and math point estimates. Among all TPPs, these estimates give a reliability of 41 percent and a heterogeneity SD of .02 (Table 1b). Among large TPPs, they give a reliability of 42 percent and a heterogeneity SD of .01 (Table 2b). These estimates are significant ($p<.05$) but still suggest little heterogeneity.

## *3.3 Missouri*

An evaluation in Missouri produced estimates for the state's "traditional, university-based" TPPs (Koedel et al., 2015). The evaluation used data from three years starting in 2008-09. It was limited to $4^{th}$-$6^{th}$ grade teachers who had no more than 4 years' experience in the fall of 2008 and who came from the 24 TPPs that contributed at least 15 teachers each to the data—1,309 teachers in 656 schools, linked to just over 61,000 student test scores in math and a similar number of test scores in "communication arts" (a synonym for ELA). A second analysis was limited to 1,000 teachers from 12 large programs that contributed at least 50 teachers each to the data.

The Missouri evaluation obtained TPP estimates from the lagged-score value-added model in (1). The left side of the model was a standardized test score in math or ELA, and the right side included a lagged test score in the same subject, along with student covariates, dummies for grade and year, and teachers' years of experience. The authors fit the model with and without school-level covariates, and they

clustered SEs at the teacher level except when they wished to illustrate the dangers of doing otherwise. Individual TPP estimates did not appear in the published article, but the authors emailed them to us.

Table 1 and Table 2 summarize Missouri TPP estimates from the model with teacher-clustered SEs and school covariates. (TPP estimates without school covariates were very similar (Koedel et al., 2015, Table 2).) Among the 12 large TPPs, *Q* tests come close to rejecting the null hypothesis of homogeneity ($.05<p<.11$), but the estimated heterogeneity SD is small in both communication arts (.01) and math (.02). Estimated reliability is about 40 percent. Among all 24 TPPs, *Q* tests convincingly reject the null hypothesis ($p<.01$), and the estimated heterogeneity SD is .03 in both math and communication arts. The estimated heterogeneity SD is larger among all TPPs than among large TPPs—but all the heterogeneity estimates are very small.

Because the estimates are noisy and the true differences are small, it is rare for a TPP to stand out as significantly different from average. After correction for multiple tests, one large TPP differ significantly from the average in math, and one small TPP differs significantly from the average in ELA.

Although Missouri's data-sharing agreement prevents us from displaying caterpillar plots of TPP estimates, we did look at caterpillar plots and they confirmed our impression that the estimates are nearly homogenous. Among large TPPs, the null distribution fits the point estimates almost perfectly in communication arts, and almost perfectly in math, except for a single TPP which, with a point estimate of -.08 SD, is significantly worse than average (after adjustment for multiple tests). Among all TPPs, the null distribution fit the math estimates very well, except for one TPP, and the null distribution fit the communication arts estimates almost perfectly through the bottom three-quarters of the distribution.
Instead of relying on SE estimates, an alternative is to estimate the correlation and the square root of the covariance between reading and math point estimates. Among all TPPs, these statistics estimate a reliability of 66 percent and a heterogeneity SD of .03 Table 1b).[10] Among large TPPs, these statistics estimate a reliability of 81 percent and a heterogeneity SD of .02 (Table 2b). These estimates are highly significant ($p<.001$) but suggest little heterogeneity.

## *3.4 New York City*

An evaluation in New York City (NYC) focused on the TPPs that "produce the vast majority of new teachers for NYC public schools" (Boyd et al., 2009), including traditional TPPs based at colleges and universities, as well as alternative route TPPs associated with Teach For America or the NYC Teaching Fellows.

To obtain TPP estimates, the NYC evaluators fit the lagged-score value-added model in (1) to the math and English language arts (ELA) scores of students taught by first and second year $4^{th}$-$8^{th}$ grade teachers in school years 2000-01 through 2005-06. Over this period, the math teachers taught 89,221 student-years in 857 schools (Boyd et al., 2009, Appendix A). The right side of the model included lagged scores in both math and ELA, squares of those scores, FEs for school, year, and grade, student demographics and prior behaviors, and classroom covariates including class size, class averages of student variables, and class averages of lagged test scores.

The NYC evaluators published math and ELA estimates for the 15 largest TPPs, which each contributed at least 50 teachers to the data, as well as math (but not ELA) estimates for all 23 TPPs, both

[10] Our reliability estimate of .66 differs just slightly from the correlation of .65 reported by the original Missouri evaluators (Koedel, Parsons, Podgursky, & Ehlert, 2015, Table 2). We believe this is because the original evaluators reported the Pearson correlation, while we report the intraclass correlation as estimated from the ANOVA model in Section 2.2.3. Both estimates are consistent for the same estimand, but can differ slightly when the number of TPPs is limited.

large and small. Table A1 in the Appendix shows the individual TPP estimates, which we copied from the NYC evaluators' Figure 1 and a table in their Appendix A (Boyd et al., 2009, pp. 428, 436).

The top of our Figure 3 shows caterpillar plots of the NYC math and ELA estimates for large TPPs. The null distribution, which is estimated from the SEs, fits the point estimates almost perfectly. Consistent with this, a *Q* test fails to reject homogeneity ($p>.6$) and the estimates of reliability and heterogeneity are zero in both math and ELA.

The bottom of Figure 3 shows results in math, but not ELA, for all TPPs, both large and small. Within this group a *Q* test rejects homogeneity ($p=.001$), and in fact the heterogeneity among TPPs appears rather substantial, with an estimated SD of .084. However, all the evidence for heterogeneity comes from a single outlier on the right. If we exclude that outlier, there is no evidence of heterogeneity; the *Q* test fails to reject homogeneity ($p>.9$), and the point estimates of reliability and heterogeneity are zero.

When results are so sensitive to a single outlier, we would like to know what the outlier really represents. Is it truly an extraordinary TPP, or is it something more disappointing, like a data error, a model specification error, or a typo in the point estimate? The NYC evaluators did not remark on the outlier, which does not stand out as clearly in their Appendix A as it does in our Figure 3. It is evidently a smaller TPP, since it does not appear in the graph of estimates from large TPPs. We would hope that the outlier represents an extraordinary TPP, but perhaps the outlier is too good to be true. It suggests that one year with a teacher from the outlying TPP, rather than an average TPP, raises math scores by over .4 SD—more than any TPP or other educational intervention that we have heard of, including 3 years in a KIPP charter school (Tuttle et al., 2013). The TPP coefficient may be overestimated, since it is highlighted conditionally on a significance test (section 2.2.2), and its SE may be underestimated, since teacher-clustered SEs are biased downward for smaller TPPs (section 2.1.1). Nevertheless, we would hesitate to either dismiss or accept a possibly extraordinary TPP without further information.

The results so far rely on SE estimates. Although concerns have been raised that the NYC evaluators did not cluster their SEs at the teacher level (Koedel et al., 2015, n. 20), our results suggest that they did. If the SEs were not teacher-clustered, the SEs would be underestimated by 50 to 150 percent (von Hippel et al., 2016, Table 2), the point estimates would be more dispersed than the null distribution, and our statistics would suggest substantial heterogeneity.[11] Instead, except for one outlier, the point estimates fit the null distribution closely, and the statistics suggest homogeneity. This suggests that the published SEs are not too small and were appropriately clustered at the teacher level. The NYC authors did not mention clustering the SEs of their TPP estimates, but they did mention teacher clustering other regressions in the same article (Boyd et al., 2009, p. 422).

Instead of relying on SE estimates, we can estimate reliability and heterogeneity from the correlation and the covariance between ELA and math point estimates for large TPPs. Among large TPPs (Table 2b), the correlation implies a reliability of 66 percent, and the square root of the covariance implies a heterogeneity SD of .03 ($p<.001$). We would like to get correlation-based estimates for all TPPs, both large and small, but that is not possible since for small TPPs we only have estimates in one subject, math.

## *3.5 Florida*

An evaluation in Florida produced estimates for 33 TPPs (Mihaly et al., 2013). These were all traditional, college-based TPPs, since although Florida has alternative-route teachers, the Florida data

[11] To put the point another way, if we believed the SEs were underestimated, we might try to correct SE estimates by multiplying them by 1.5 to 2.5, but this would yield a null distribution that was substantially *more* dispersed than the point estimates—an unlikely result.

lumps them together and does not assign them to individual TPPs. The evaluation focused on inexperienced teachers, defined as teachers in their first, second, or third year of teaching. While the largest TPP produced 496 inexperienced teachers over the five years of the study, many of the TPPs were small: half produced fewer than 40 inexperienced teachers each, 5 TPPs produced fewer than 5 inexperienced teachers each, and one TPP produced only one teacher.

The evaluators' value-added model used 4th and 5th grade test scores from a five-year period from 2000-2001 through 2004-2005. The scores came from the multi-subject Florida Comprehensive Assessment Test; the evaluators used the overall score and did not analyze scores for component subjects such as reading and math. The evaluators fit a lagged-score value-added model (1) whose control variables included student demographics, teacher experience, grade dummies, and year dummies.

The evaluators fit four versions of the model: one with school FEs, one with a reference group of experienced teachers, one with both, and one with neither. As stated in the Methods section, these decisions change what quantity is being estimated. In particular, with school FEs and experienced teachers, we are not just comparing TPPs to each other; we are also comparing their recent graduates to the experienced teachers in the schools where they are hired.[12] Therefore we expect the combination of school FEs and an experienced reference group to result in larger estimates of heterogeneity.

We copied TPP estimates and SEs from the Florida evaluators' Appendix Tables A.1 and A.2 (Mihaly et al., 2013). Our results include all TPPs, and we also present results for large TPPs that contributed at least 40 teachers to the data.

The Florida evaluators did not cluster their SEs,[13] so their reported SEs were underestimated. Teacher-clustered SEs can exceed unclustered SEs by 50 to 150 percent (von Hippel et al., 2016, Table 2). To approximate the effect of teacher clustering in Florida, our main results take a middle road and inflate the reported SEs by 100 percent. In the Appendix, Table A2 presents alternative Florida SEs inflated by 50 or 150 percent.

The Florida TPP estimates are summarized in Figure 4-Figure 5, and Table 1-Table 2. The results are sensitive to the sample and model. The estimated heterogeneity SD is zero when the sample is restricted to large TPPs, but larger (.04-.08) when the sample includes all TPPs. But remember that the all-TPP sample includes six very small TPPs with just one to five teachers in the data.

Among all TPPs, the largest heterogeneity estimate comes from the model that combines school FEs with a comparison group of experienced teachers. This model probably overestimates heterogeneity, since heterogeneity among the TPPs is confounded with heterogeneity in the comparison group (see Section 2.1.3). Among large TPPs, the choice of model has less effect on the results.

In the Appendix, Table A2 presents alternative Florida estimates with SEs inflated by a smaller amount (50 percent) or by a larger amount (150 percent). Naturally, more inflated SEs result in lower estimates for heterogeneity, reliability, and the number of TPPs that differ significantly from the mean. Less inflated SEs have the opposite effect.

Instead of relying on the SE estimates, an alternative is to estimate reliability and heterogeneity using the correlation and the square root of the covariance between estimates from tests of different subjects. Unfortunately, this is not possible in the Florida results, which do not report results separately by subject.

[12] The evaluation also includes large comparison groups of inexperienced teachers with alternate or out-of-state certifications.

[13] The Florida evaluation does not mention clustering the SEs. The corresponding author confirmed that SEs were not clustered (Kata Mihaly, email communication, April 10, 2017).

## *3.6 Louisiana*

An evaluation in Louisiana produced estimates for 10 traditional and alternative TPPs (Gansle et al., 2012). The evaluation drew data from 3 school years and was restricted to TPPs with data on at least 25 new (first year) 4th-9th grade teachers in a given subject. By that criterion, ten TPPs, with 25-68 ELA teachers each, had estimates in ELA; 8 TPPs had estimates in math; and 7 TPPs had estimates in reading, science, and social studies. Note that the Louisiana data has fewer teachers and fewer TPPs than the data from other states.

From the Louisiana results, we copied TPP point estimates along with the Louisiana evaluators' "68% CIs," which imply the SE since a 68% CI extends one SE in each direction. Both the point estimates and the SEs were reported to 2 significant digits on the metric of Louisiana's state tests, which had an SD of 5. We standardized the results to a scale where the tests have an SD of 1.

The Louisiana evaluation fit the lagged-score value-added model in (1). The left side was a test score in one subject, and the right side included lagged test scores in all subjects; student, classroom, teacher, and school district covariates; and REs at the school and classroom level (Gansle et al., 2012, Appendix Table I).[14]

Louisiana's use of classroom REs rather than teacher REs yields SE estimates that are too small; SE estimates are 18 percent smaller with classroom REs than with teacher REs, according to estimate rom Missouri teacher REs.[15] To correct this, we added 18 percent to the reported SEs in Louisiana. Table A3 in the Appendix shows what our results would look like with the original SEs.

Louisiana's estimation sample included a comparison group of experienced teachers. As discussed earlier, in a model with school REs, the use of an experienced comparison group tends to increase estimates of heterogeneity, though not as much as it does with school FEs (Sections 2.1.3 and 3.5).

Figure 6 summarizes the TPP estimates in Louisiana. In four of the five subjects—ELA, reading, science, and social studies—little heterogeneity is evident. The *Q* tests are nonsignificant ($p>.1$), and no individual TPPs differ significantly from the average. This is true both among large TPPs and among all TPPs.

In the fifth subject, math, some heterogeneity is evident. The *Q* test is significant ($p<.03$), both among large TPPs and among all TPPs. Figure 6 shows that the math heterogeneity comes entirely from one large, above-average TPP. This TPP's contrast sits well above the null distribution and has a 95% Bonferroni CI that does not reach zero. This TPP is identified in the Louisiana analysis as "private practitioner TPP 2," an alternative certification program that was not based at a college or university (Gansle et al., 2012). Its estimated effect size is .14 SD, which is neither trivially small nor implausibly large; in fact, it is similar to math effects in randomized controlled trials of the alternative TPP Teach for America (Clark et al., 2013; Clark, Isenberg, Liu, Makowsky, & Zukiewicz, 2017; Decker, Mayer, & Glazerman, 2004).

---

[14] The Louisiana evaluators describe their model as having three levels: the student level, the "teacher/classroom" level, and the school level. The idea of a "teacher/classroom level" is ambiguous, since a teacher can have more than one classroom. Inspecting the details of the Louisiana model makes it clear that, while the *regressors* come from both the teacher level and the classroom level, the RE is unambiguously at the classroom level (Gansle, Noell, & Burns, 2012, Appendix Table I).

[15] We calculated this 18 percent figure by comparing teacher- and classroom-clustered SEs shared with us by the Missouri evaluators. While these SEs have not appeared in print, one can get a similar figure by comparing the "estimation-error variance share" that the Missouri evaluators report for teacher- and classroom-clustered models (Koedel et al., 2015, Table 4). The connection is that the estimation-error variance share is proportional to the mean of the squared SEs. See our equation (2).

Table A3 in the Appendix shows how our Louisiana estimates change if we do not inflate the reported SEs. Estimates of heterogeneity are larger, but only slightly. There is still only one TPP that differs significantly from the average in math, and none that differ significantly from average in other subjects.

Instead of relying on the SE estimates, an alternative is to estimate the correlation and the square root of the covariance between estimates from tests of different subjects. Since 5 subjects were tested in the Louisiana evaluation, the correlation generalizes to the intraclass correlation and the square root of the covariance generalizes to the SD between TPPs (Section 2.2.3). Among all TPPs, these estimates give a reliability of 55 percent and a heterogeneity SD of .04 (Table 1b). Among large TPPs, they give a reliability of 77 percent and a heterogeneity SD of .05 (Table 2b). These estimates are significant ($p<.05$) and suggest about the same amount of heterogeneity as we estimated by comparing point estimates to SEs.

# 4 Results across all states

Table 1 and Table 2 summarize the results across all states.

The results suggest little heterogeneity. When we compare point estimates to SEs, the average heterogeneity SD is .03 among all TPPs (Table 1a), and shrinks to just .01 if we limit the estimates to large TPPs (Table 2a).

This last result, taken at face value, suggests that there is more heterogeneity among small TPPs than among large TPPs. This makes sense if small TPPs rely on exceptional individuals or resource-intensive techniques that are hard to scale. It also makes sense if, like large restaurant chains, large TPPs must standardize practices in ways that produce consistent but unexceptional results.

On the other hand, the heterogeneity among small TPPs may be exaggerated. When estimated using SEs, heterogeneity will be overestimated if the SEs are underestimated—as they are when teacher clustering is used in small TPPs. In fact, if we ignore the SEs and simply compare point estimates across different subjects, we find that heterogeneity estimates are similar for small and large TPPs. The average heterogeneity estimate is .03 SD among all TPPs (Table 1b), and also .03 SD among large TPPs (Table 2b).

In any case, the heterogeneity is very small, and the heterogeneity estimates are somewhat uncertain. We should not make too much of the difference between estimates of .01 and .03 SD.

Despite the low heterogeneity, we can occasionally single out a TPP as significantly better or worse than average. Here it is important to correct for multiple tests. If we neglected the multiple tests issue, and simply conducted individual hypothesis tests with a significance level of .05, we would flag 7-9% of TPPs as significantly different—but 5% of differences would be expected to be significant by chance alone. If we correct for multiple comparisons, only 2-3% of TPPs stand out as significantly different overall, and within each state and subject at most one TPP differs significantly from the average—typically by .05 SD or less. An exception occurs in Louisiana, where one large TPP differs by .15 SD from the average, but only in math. There is also an outlier in NYC, but it is a small TPP, and its estimate seems implausibly large (as discussed in Section 3.4).

# 5 Conclusion

Before we conducted these reanalyses, it seemed to us that evaluations had yielded discrepant estimates in different states. In Missouri, Texas, and Washington state, the differences between TPPs appeared small or negligible, whereas in Louisiana and NYC the differences appeared more substantial. In our reanalyses, though, the differences were small even in Louisiana and NYC. In NYC, there was no

heterogeneity among large TPPs, and no heterogeneity among small TPPs except for a single outlier. In Louisiana, the heterogeneity was nonsignificant in 4 out of 5 subjects. Across all states, the average heterogeneity between TPPs is .01 to .03 SD—just as we predicted from the back-of-the-envelope calculation in our Introduction.

A variety of practices can make heterogeneity appear greater than it is. SEs may be underestimated. Results may be graphed using short CIs that extend only one SE in each direction. Even 95 percent CIs that extend 2 SEs in each direction may still be too short if they fail to correct for multiple tests. When experienced teachers are used as a comparison group, the variance among experienced teachers may be confounded with the variance among TPPs. Each of these issues has occurred in at least one TPP evaluation. We have tried to compensate for them in our review.

A caterpillar plot of point estimates may look deceptively heterogeneous if it is not compared to a null distribution showing what the point estimates would look like if the TPPs were identical. In our results, we have graphed the null distribution, and we have released a new Stata command—*caterpillar*—which will help others to graph the null distribution as well.

Because there is little heterogeneity and substantial estimation error, it is rarely possible to single out a large TPP that is significantly better or worse than average. If policy decisions are made on the basis of TPPs that have been mistakenly labeled as exceptional, those decisions may be ineffective or even counterproductive.

Despite the danger of singling out a TPP erroneously, if we are careful, we can occasionally identify a TPP that may be truly exceptional. In every state except for NYC, there was one large TPP that appeared to be significantly different from average in at least one subject, even after adjustment to the SEs and correction for multiple tests. In Florida, Missouri, Texas, and Washington, the exceptional TPPs had very small effects (.03-.05 SD), but in Louisiana, the effect of the exceptional TPP on math scores was somewhat larger (.15 SD).

In short, TPP evaluations may have some policy value, but the value is more limited than was originally envisioned. It is not meaningful to rank all the TPPs in a state. The true differences between most TPPs are too small to matter, and the estimated differences consist mostly of noise. Nevertheless, in some states and subjects it is possible to single out an individual TPP whose teachers may really be better or worse than average. Perhaps we can learn something about TPP effectiveness by further studying individual TPPs that stand out in a value-added evaluation.

## References


115th Congress. Providing for congressional disapproval under chapter 8 of title 5, United States Code, of the rule submitted by the Department of Education relating to teacher preparation issues, Pub. L. No. House Joint Resolution 58 (2017).

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.

Burnham, K. P., & Anderson, D. R. (1998). *Model selection and inference: a practical information-theoretic approach*. New York: Springer.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood. *American Economic Review*, *104*(9), 2633–2679. https://doi.org/10.1257/aer.104.9.2633

Clark, M. A., Chiang, H. S., Silva, T., McConnell, S., Sonnenfeld, K., Erbe, A., & Puma, M. (2013). *The Effectiveness of Secondary Math Teachers from Teach For America and the Teaching Fellows*

*Programs* (No. NCEE 2013-4015). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences.

Clark, M. A., Isenberg, E., Liu, A. Y., Makowsky, L., & Zukiewicz, M. (2017). *Impacts of the Teach For America Investing in Innovation Scale-Up* (No. 06889.740). Princeton, NJ: Mathematica Policy Research. Retrieved from https://www.mathematica-mpr.com/our-publications-and-findings/publications/impacts-of-the-teach-for-america-investing-in-innovation-scaleup

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

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

Decker, P. T., Mayer, D. P., & Glazerman, S. (2004). *The effects of Teach for America on students: Findings from a national evaluation*. Mathematica Policy Research.

Department of Education, O. of P. E. Teacher Preparation Issues, Pub. L. No. Federal Register, 81 FR 75494, 75494 (2016). Retrieved from https://www.federalregister.gov/documents/2016/10/31/2016-24856/teacher-preparation-issues

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. https://doi.org/10.1177/0022487112439894

Gelman, A. (2017). The Failure of Null Hypothesis Significance Testing When Studying Incremental Changes, and What to Do About It. *Personality and Social Psychology Bulletin*, 0146167217729162. https://doi.org/10.1177/0146167217729162

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

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

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

Koedel, C. (2008). Teacher quality and dropout outcomes in a large, urban school district. *Journal of Urban Economics*, *64*(3), 560–572. https://doi.org/10.1016/j.jue.2008.06.004

Koedel, C., & Parsons, E. (2014). Evaluating Teacher Preparation Programs Using the Performance of their Graduates. Retrieved from http://www.tcrecord.org/Content.asp?ContentId=17741

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. https://doi.org/10.1162/EDFP_a_00172

Koretz, D. M. (2002). Limitations in the Use of Achievement Tests as Measures of Educators' Productivity. *The Journal of Human Resources*, *37*(4), 752–777. https://doi.org/10.2307/3069616

Koretz, D. M. (2009). *Measuring Up: What Educational Testing Really Tells Us*. Cambridge, Mass.: Harvard University Press.

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

MacKinnon, J. G., & Webb, M. D. (2017). The wild bootstrap for few (treated) clusters. *The Econometrics Journal*, n/a-n/a. https://doi.org/10.1111/ectj.12107

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.

Quezada-Hofflinger, A., & von Hippel, P. T. (2017). *The Response to High Stakes Testing in Chile, 2005-2013: Legitimate and Illegitimate Ways to Raise Test Scores* (SSRN Scholarly Paper No. ID 2906552). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=2906552

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

Ronfeldt, M., & Campbell, S. L. (2016). Evaluating Teacher Preparation Using Graduates' Observational Ratings. *Educational Evaluation and Policy Analysis*, 0162373716649690. https://doi.org/10.3102/0162373716649690

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. https://doi.org/10.1016/0272-7757(93)90040-N

Steinberg, M. P., & Garrett, R. (2016). Classroom Composition and Measured Teacher Performance: What Do Teacher Observation Scores Really Measure? *Educational Evaluation and Policy Analysis*, *38*(2), 293–317. https://doi.org/10.3102/0162373715616249

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

Tuttle, C. C., Gill, B., Gleason, P., Knechtel, V., Nichols-Barrer, I., & Resch, A. (2013). *KIPP Middle Schools: Impacts on Achievement and Other Outcomes. Final Report*. Mathematica Policy Research, Inc. Retrieved from http://eric.ed.gov/?id=ED540912

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 I^2 can be biased in small meta-analyses. *BMC Medical Research Methodology*, *15*(1), 35.

von Hippel, P. T., Bellows, L., Osborne, C., Lincove, J. A., & Mills, N. (2016). Teacher quality differences between teacher preparation programs: How big? How reliable? Which programs are different? *Economics of Education Review*, *53*, 31–45. https://doi.org/10.1016/j.econedurev.2016.05.002

Walsh, K. (2001). *Teacher Certification Reconsidered: Stumbling for Quality*. Baltimore, MD: The Abell Foundation.

Whitehurst, G. J., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating Teachers with Classroom Observations: Lessons Learned in Four Districts*. Brookings Institution.

A revision of this article was published in *Economics of Education Review*, *64*, 298-312.

# Tables

A revision of this article was published in *Economics of Education Review*, *64*, 298-312.

Table 1. All TPPs: Summary of estimates

a. Comparing point estimates to SEs

| State | Subject | Experienced Teacher Comparison? | School Fixed Effects? | TPPs | Homogeneity test Q | df | p | Heterogeneity SD | Reliability | Significantly different TPPs Uncorrected | Bonferroni corrected |
|---|---|---|---|---|---|---|---|---|---|---|---|
| TX | Math | No | No | 95 | 122 | 94 | .03 | .05 | .23 | 7 (7%) | 0 (0%) |
| | Reading | No | No | 92 | 121 | 91 | .02 | .02 | .25 | 7 (8%) | 0 (0%) |
| WA | Math | No | No | 20 | 19 | 19 | .44 | .00 | .02 | 0 (0%) | 0 (0%) |
| | Reading | No | No | 20 | 47 | 19 | .00 | .02 | .60 | 4 (20%) | 1 (5%) |
| MO | Math | No | No | 24 | 44 | 23 | .01 | .03 | .47 | 3 (13%) | 1 (4%) |
| | Communication arts | No | No | 24 | 47 | 23 | <.01 | .03 | .51 | 4 (17%) | 1 (4%) |
| NYC | Math, with outlier | No | No | 23 | 50 | 22 | <.01 | .08 | .56 | 1 (4%) | 1 (5%) |
| | Math, without outlier | No | No | 23 | 13 | 21 | .92 | .00 | .00 | 0 (0%) | 0 (0%) |
| FL | Composite | No | No | 33 | 64 | 32 | <.01 | .04 | .50 | 4 (12%) | 0 (0%) |
| | | Yes | No | 33 | 69 | 32 | <.01 | .06 | .54 | 4 (12%) | 1 (3%) |
| | | No | Yes | 33 | 37 | 32 | .24 | .05 | .14 | 3 (9%) | 0 (0%) |
| | | Yes | Yes | 33 | 52 | 32 | .01 | .08 | .39 | 5 (15%) | 0 (0%) |
| LA | Math | Yes | No | 8 | 16 | 7 | .03 | .04 | .55 | 1 (13%) | 1 (14%) |
| | Reading | Yes | No | 7 | 10 | 6 | .13 | .04 | .39 | 0 (0%) | 0 (0%) |
| | ELA | Yes | No | 10 | 13 | 9 | .15 | .03 | .33 | 0 (0%) | 0 (0%) |
| | Science | Yes | No | 7 | 7 | 6 | .35 | .01 | .10 | 0 (0%) | 0 (0%) |
| | Social Studies | Yes | No | 7 | 3 | 6 | .75 | .00 | .00 | 0 (0%) | 0 (0%) |
| | | | | | All estimates | | Mean | .03 | .33 | 8% | 2% |
| | | | Estimates without experienced teachers or school FEs | | | | Mean | .03 | .35 | 9% | 2% |

b. Comparing point estimates across subjects

| State | # TPPs | Heterogeneity SD | Reliability | P |
|---|---|---|---|---|
| TX | 87 | .04 | .42 | <.01 |
| WA | 20 | .02 | .41 | .03 |
| MO | 24 | .03 | .66 | <.01 |
| LA | 8 | .04 | .55 | <.01 |
| Mean | | .03 | .51 | |

Table 2. Large TPPs: Summary of estimates

a. Comparing point estimates to SEs

| State | Subject | Experienced Teacher Comparison? | School Fixed Effects? | TPPs | Homogeneity test Q | df | p | Heterogeneity SD | Reliability | Significantly different TPPs Uncorrected | Bonferroni corrected |
|---|---|---|---|---|---|---|---|---|---|---|---|
| TX | Math | No | No | 48 | 60 | 47 | .09 | .01 | .22 | 2 (4%) | 0 (0%) |
| | Reading | No | No | 37 | 50 | 36 | .06 | .00 | .28 | 3 (8%) | 1 (3%) |
| WA | Math | No | No | 18 | 18 | 17 | .42 | .01 | .03 | 0 (0%) | 0 (0%) |
| | Reading | No | No | 18 | 43 | 17 | <.01 | .02 | .60 | 3 (17%) | 1 (6%) |
| MO | Math | No | No | 12 | 19 | 11 | .06 | .02 | .43 | 2 (17%) | 1 (8%) |
| | Communication arts | No | No | 12 | 17 | 11 | .10 | .01 | .36 | 2 (17%) | 0 (0%) |
| NYC | Math | No | No | 15 | 9 | 14 | .86 | .00 | .00 | 0 (0%) | 0 (0%) |
| | ELA | No | No | 15 | 11 | 14 | .66 | .00 | .00 | 1 (7%) | 0 (0%) |
| FL | Composite | No | No | 16 | 28 | 15 | .02 | .00 | .47 | 2 (13%) | 1 (6%) |
| | | Yes | No | 16 | 33 | 15 | .01 | .00 | .54 | 3 (19%) | 1 (6%) |
| | | No | Yes | 16 | 12 | 15 | .71 | .00 | .00 | 0 (0%) | 0 (0%) |
| | | Yes | Yes | 16 | 19 | 15 | .21 | .00 | .21 | 1 (6%) | 0 (0%) |
| LA | Math | Yes | No | 7 | 15 | 6 | .02 | .05 | .61 | 1 (14%) | 1 (14%) |
| | Reading | Yes | No | 3 | 1 | 2 | .54 | .00 | .00 | 0 (0%) | 0 (0%) |
| | ELA | Yes | No | 6 | 8 | 5 | .14 | .04 | .40 | 0 (0%) | 0 (0%) |
| | Science | Yes | No | 2 | 0 | 1 | .66 | .00 | .00 | 0 (0%) | 0 (0%) |
| | Social Studies | Yes | No | 4 | 3 | 3 | .46 | .00 | .00 | 0 (0%) | 0 (0%) |
| | | | | | All estimates | | Mean | .01 | .24 | 7% | 3% |
| | | Estimates without experienced teachers or school FEs | | | | | Mean | .01 | .27 | 9% | 3% |

b. Comparing point estimates across subjects

| State | # TPPs | Heterogeneity SD | Reliability | p |
|---|---|---|---|---|
| TX | 36 | .01 | .30 | .05 |
| WA | 18 | .01 | .42 | .03 |
| MO | 12 | .02 | .81 | <.01 |
| NYC | 15 | .03 | .66 | <.01 |
| LA | 7 | .05 | .77 | <.01 |
| Mean | | .03 | .59 | |

# Figures

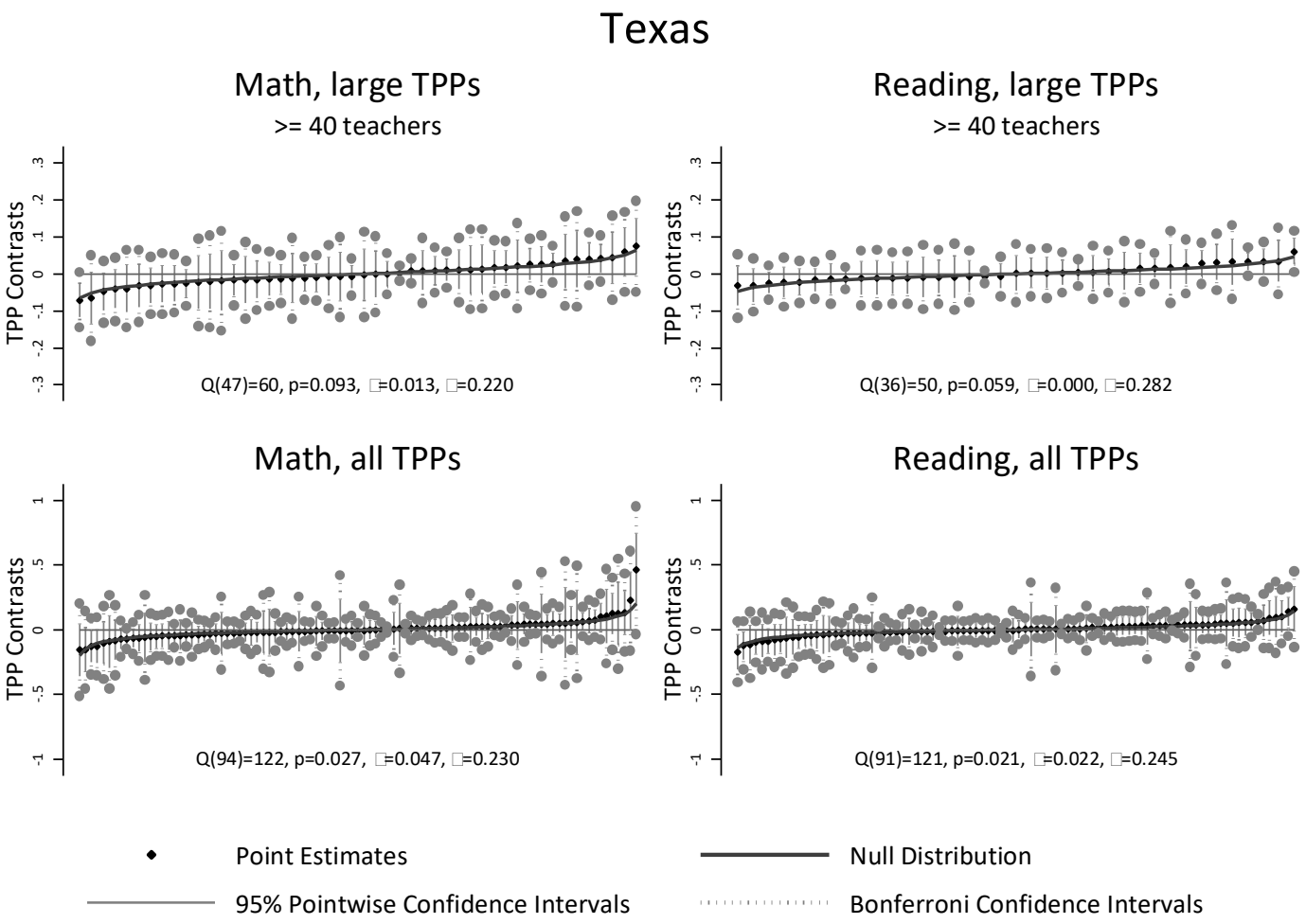


Figure 1. TPP contrasts in Texas.

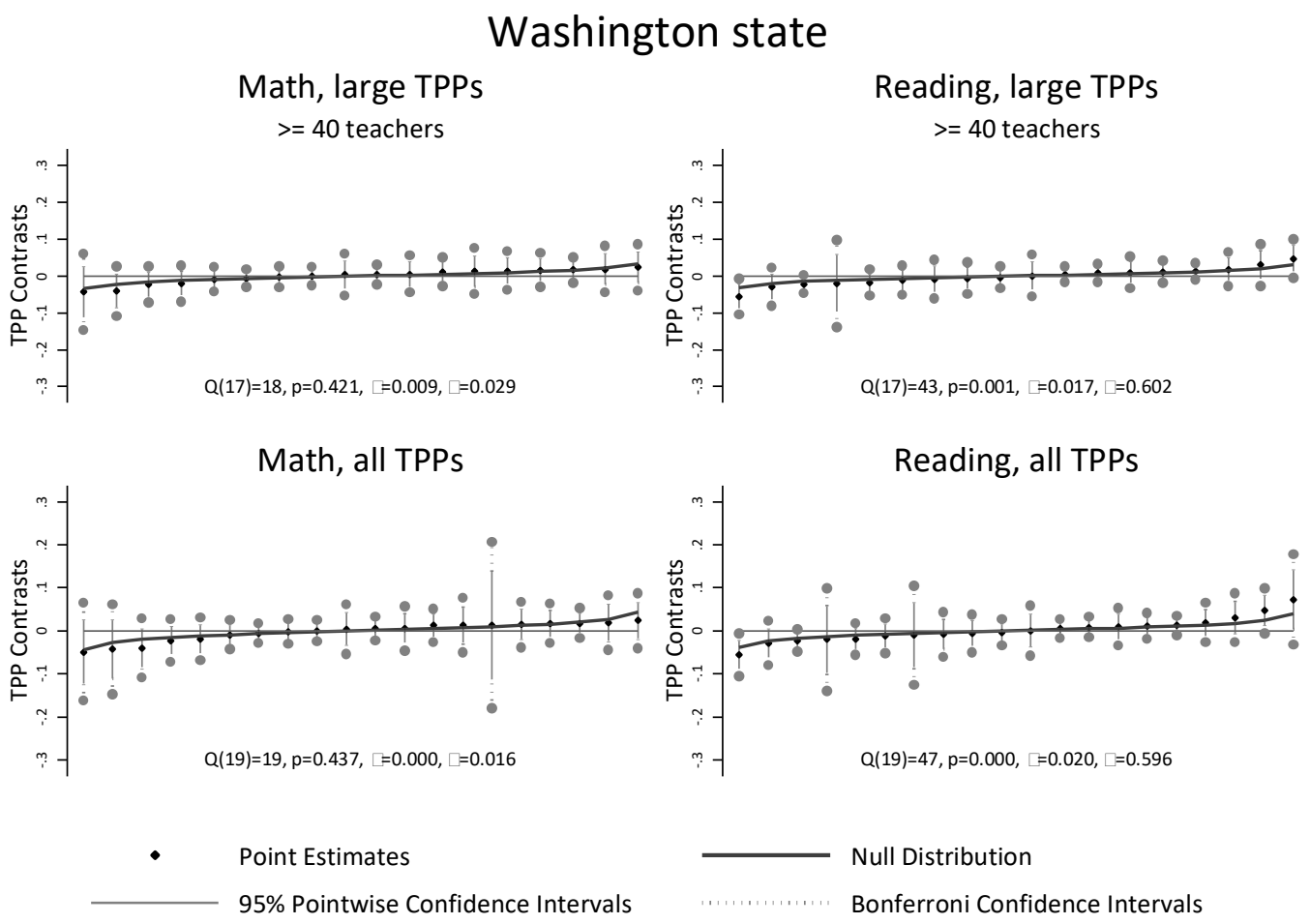


Figure 2. TPP contrasts in Washington state. There is heterogeneity in reading but not in math. In reading, after adjustment for multiple tests, only the leftmost TPP is significantly worse than average.

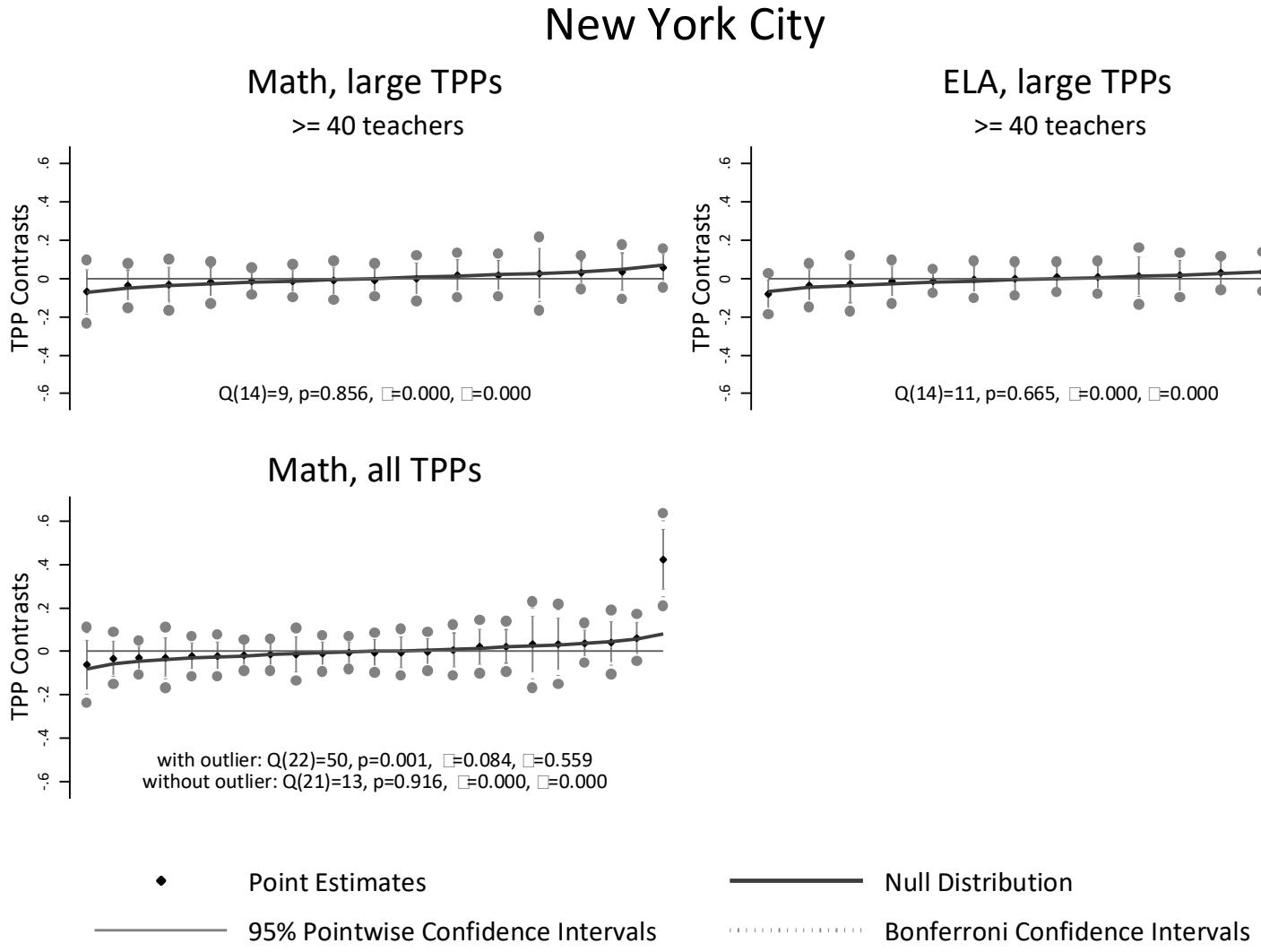


Figure 3. TPP estimates in New York City. There is no evidence of heterogeneity except for one small outlying TPP in math.

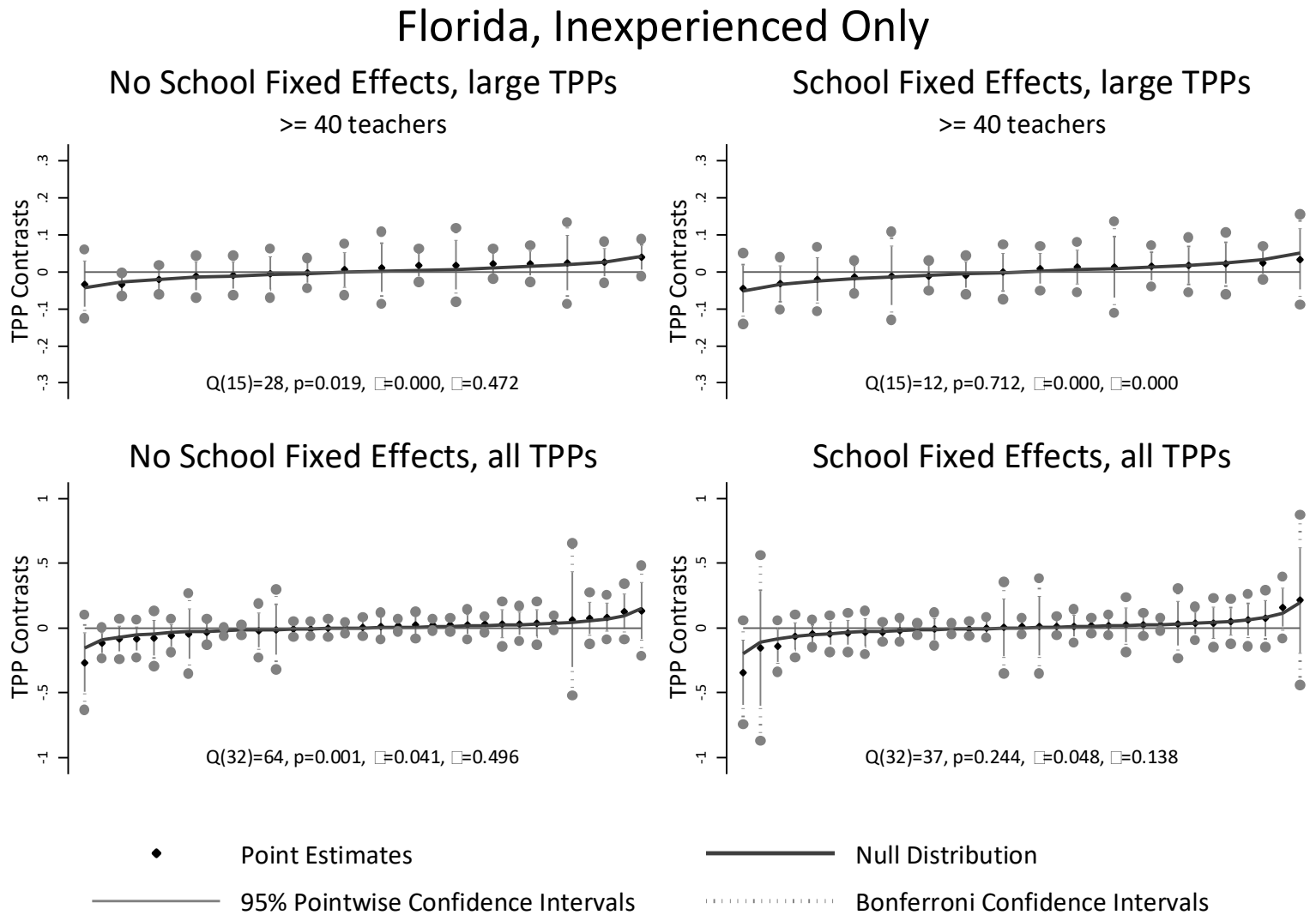


Figure 4. Florida estimates from models fit to inexperienced teachers only.

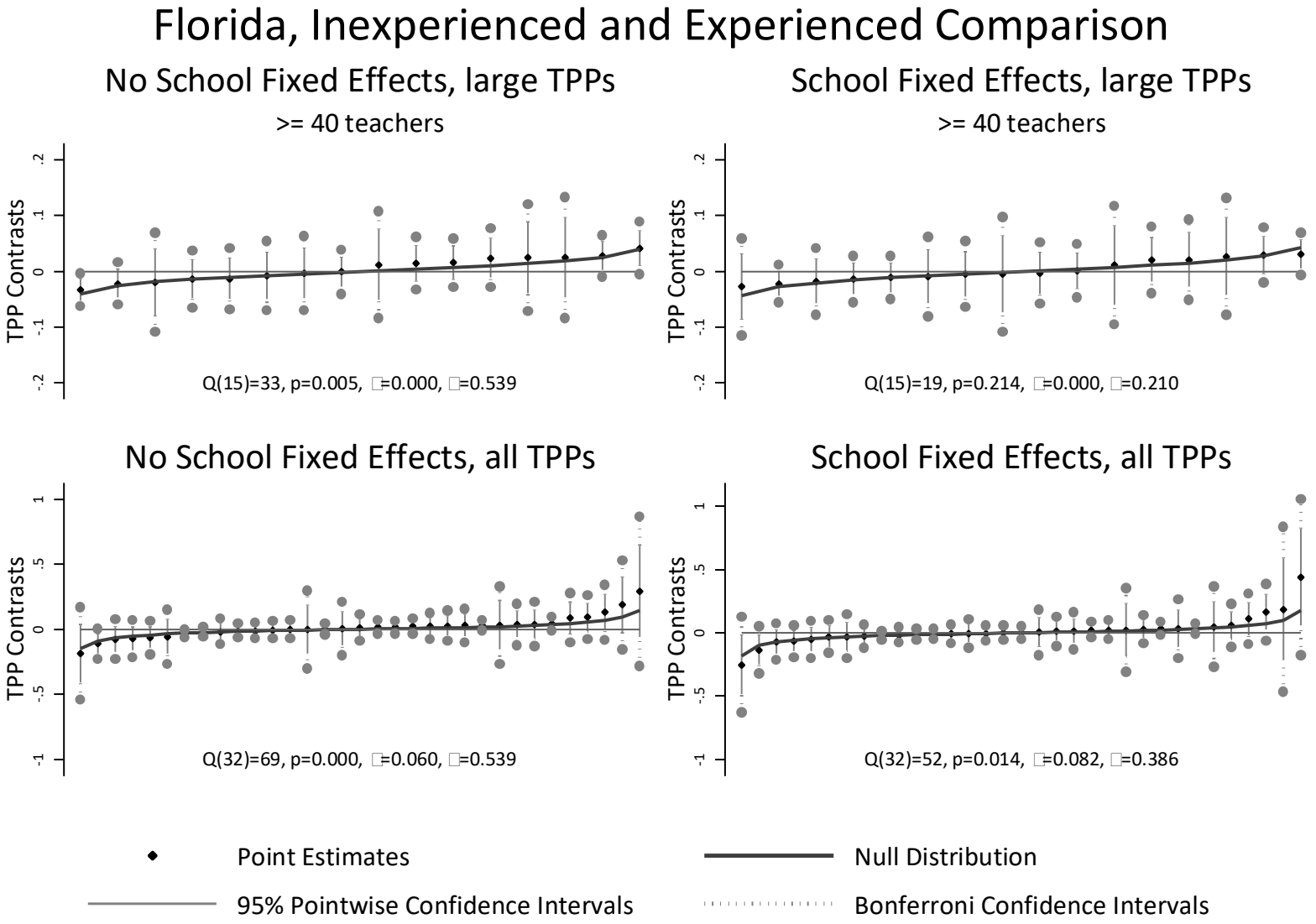


Figure 5. Florida estimates from models fit to inexperienced teachers and an experienced comparison group.

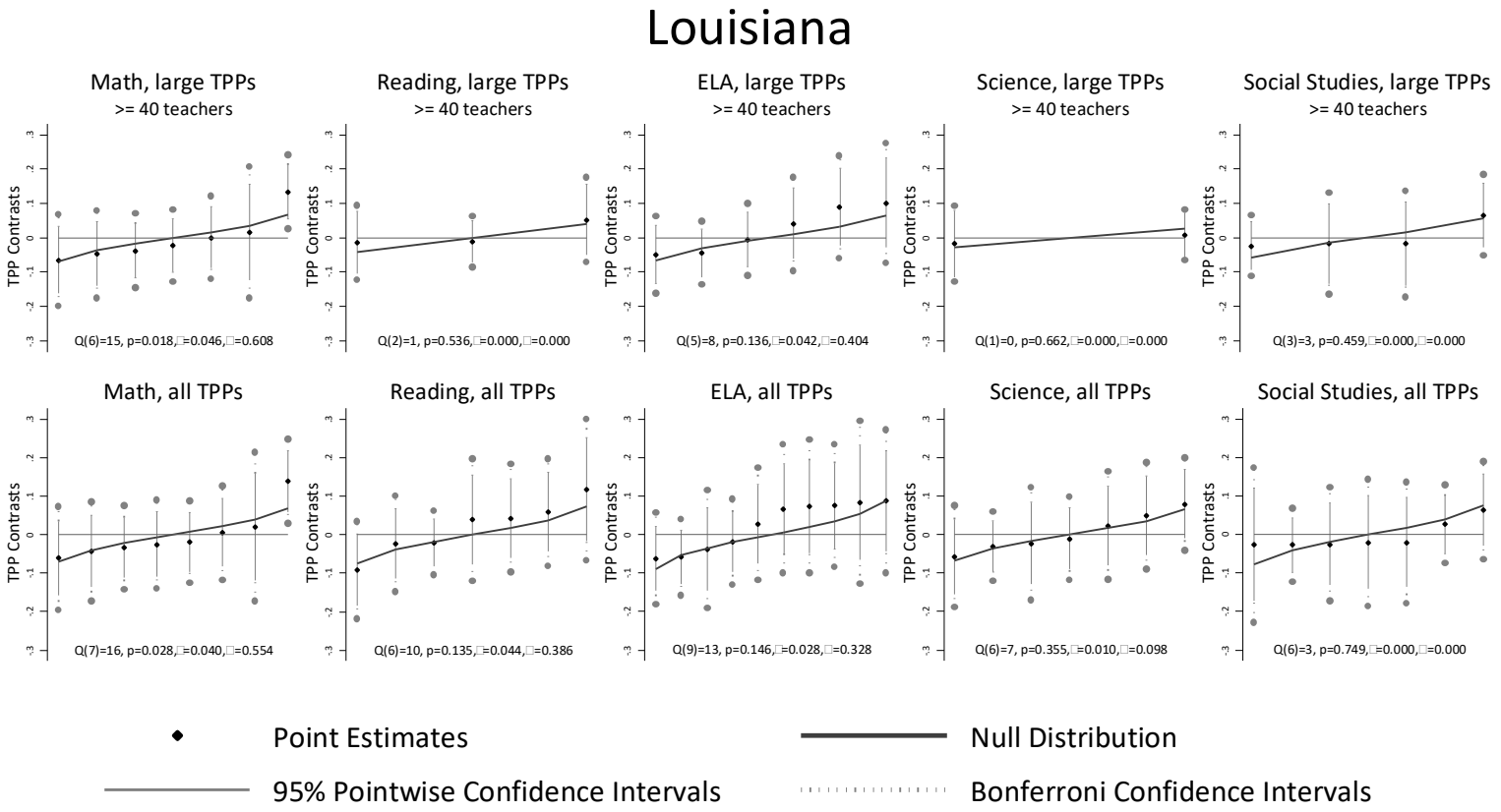


Figure 6. TPP contrasts in Louisiana. There is no evidence for heterogeneity except for one large TPP in math.

A revision of this article was published in *Economics of Education Review, 64*, 298-312.

# Appendix

Table A1. Individual TPP estimates from the NYC evaluation

| From Figure 1 | | | | | From Appendix A | | |
|---|---|---|---|---|---|---|---|
| | ELA | | Math | | | Math | |
| TPP | Est | SE | Est | SE | TPP | Est | SE |
| | | | | | 0 | -.007 | (.030) |
| 1 | .058 | (.035) | .076 | (.034) | 1 | .078 | (.035) |
| 2 | -.004 | (.049) | -.048 | (.056) | 2 | -.047 | (.057) |
| | | | | | 3 | .049 | (.060) |
| 3 | .020 | (.031) | .052 | (.029) | 4 | .054 | (.030) |
| 4 | .063 | (.039) | .020 | (.041) | 5 | .022 | (.039) |
| 5 | .006 | (.038) | .001 | (.037) | 6 | .002 | (.040) |
| 6 | .029 | (.029) | .009 | (.035) | 7 | .011 | (.035) |
| 7 | .080 | (.055) | .047 | (.066) | 8 | .048 | (.065) |
| 8 | .017 | (.033) | .037 | (.038) | 9 | .039 | (.038) |
| 9 | .050 | (.030) | .008 | (.029) | 10 | .010 | (.030) |
| 10 | -.013 | (.039) | -.012 | (.046) | 11 | -.012 | (.045) |
| 11 | -.059 | (.036) | -.017 | (.039) | 12 | -.017 | (.040) |
| 12 | .037 | (.039) | .037 | (.040) | 13 | .038 | (.040) |
| 13 | .029 | (.027) | .012 | (.029) | 14 | .014 | (.029) |
| 14 | .033 | (.051) | .056 | (.048) | 15 | .058 | (.048) |
| 15 | .007 | (.022) | .007 | (.024) | 16 | .009 | (.024) |
| | | | | | 17 | .440 | (.070) |
| | | | | | 18 | .001 | (.024) |
| | | | | | 19 | -.014 | (.026) |
| | | | | | 20 | -.005 | (.032) |
| | | | | | 21 | -.001 | (.023) |
| | | | | | 22 | .006 | (.027) |

*Note*. Estimates copied from Figure 1 and Appendix A in the NYC evaluation (Boyd et al., 2009). The NYC evaluators' Figure 1 gave math and ELA estimates for the 15 largest TPPs, while their Appendix A included all TPPs but only reported estimates for math. Note that the TPP identification numbers differ slightly between their Figure 1 and their Appendix A; TPPs 1 and 2 in their Figure 1 match TPPs 1 and 2 in their Appendix A, while TPPs 3 to 15 in their Figure 1 match TPPs 4 to 16 in their Appendix A. In math, their Figure 1 and Appendix A agree perfectly to 2 decimal places; minor disagreements in the third decimal place are likely to measurement error in our copying estimates from their Figure 1, which we did using an on-screen ruler in a PDF reader. Note that their Appendix A gave point estimates and *t* statistics, which we converted to SEs. Their Figure 1 gave point estimates and "standard error ranges," defined as intervals extending one (not two) SE above and below each point estimate. This definition of the standard error range as covering 1 SE in each direction is implicit in the NYC article's statement that "Figure 1…shows the standard error of each estimate," and "there is more than a two-standard error difference between the higher and lower value-added programs" (Boyd et al., 2009, pp. 428–429). If the standard error range covered 2 SEs rather than 1 in each direction, our Table A1 would not show such strong agreement between their Figure 1 and their Appendix A.

A revision of this article was published in *Economics of Education Review*, *64*, 298-312.

Table A2. Summary of Florida estimates, with SEs inflated by different amounts

| | Experienced teacher comparison? | School FEs? | SE inflation | # TPPs | *Q* | *df* | *p* | Heterogeneity SD | Reliability | Significantly different TPPs | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | Uncorrected | Bonferroni corrected |
| All TPPs | Yes | Yes | none | 33 | 209 | 32 | <.001 | .10 | .85 | 13 | 9 |
| | | | 50% | | 93 | 32 | <.001 | .09 | .66 | 10 | 1 |
| | | | 100% | | 52 | 32 | .014 | .08 | .39 | 5 | 0 |
| | | | 150% | | 33 | 32 | .400 | .06 | .04 | 1 | 0 |
| | | No | none | | 278 | 32 | <.001 | .08 | .89 | 17 | 13 |
| | | | 50% | | 123 | 32 | <.001 | .07 | .74 | 13 | 3 |
| | | | 100% | | 69 | 32 | <.001 | .06 | .54 | 4 | 1 |
| | | | 150% | | 44 | 32 | .071 | .04 | .28 | 3 | 0 |
| | No | Yes | none | | 149 | 32 | <.001 | .08 | .79 | 10 | 4 |
| | | | 50% | | 66 | 32 | <.001 | .07 | .52 | 4 | 1 |
| | | | 100% | | 37 | 32 | .244 | .05 | .14 | 3 | 0 |
| | | | 150% | | 24 | 32 | .853 | .00 | .00 | 1 | 0 |
| | | No | none | | 254 | 32 | <.001 | .07 | .87 | 18 | 8 |
| | | | 50% | | 113 | 32 | <.001 | .06 | .72 | 11 | 3 |
| | | | 100% | | 64 | 32 | .001 | .04 | .50 | 4 | 0 |
| | | | 150% | | 41 | 32 | 0.14 | 0 | 0.21 | 2 | 0 |
| Large TPPs | Yes | Yes | none | 16 | 76 | 15 | <.001 | .02 | .80 | 4 | 3 |
| | | | 50% | | 34 | 15 | .004 | .01 | .56 | 3 | 1 |
| | | | 100% | | 19 | 15 | .214 | .00 | .21 | 1 | 0 |
| | | | 150% | | 12 | 15 | .667 | .00 | .00 | 1 | 0 |
| | | No | none | | 130 | 15 | <.001 | .02 | .89 | 6 | 4 |
| | | | 50% | | 58 | 15 | <.001 | .01 | .74 | 4 | 2 |
| | | | 100% | | 33 | 15 | .005 | .00 | .54 | 3 | 1 |
| | | | 150% | | 21 | 15 | .143 | .00 | .28 | 2 | 0 |
| | No | Yes | none | | 46 | 15 | <.001 | .02 | .68 | 3 | 1 |
| | | | 50% | | 21 | 15 | .152 | .01 | .27 | 1 | 0 |
| | | | 100% | | 12 | 15 | .712 | .00 | .00 | 0 | 0 |
| | | | 150% | | 7 | 15 | .946 | .00 | .00 | 0 | 0 |
| | | No | none | | 114 | 15 | <.001 | .02 | .87 | 8 | 4 |
| | | | 50% | | 50 | 15 | <.001 | .01 | .70 | 4 | 2 |
| | | | 100% | | 28 | 15 | .019 | .00 | .47 | 2 | 1 |
| | | | 150% | | 18 | 15 | .254 | .00 | .18 | 1 | 0 |

Table A3. Summary of Louisiana estimates, with SEs either inflated or not

| | | | | | | | | | Significantly different TPPs | |
|---|---|---|---|---|---|---|---|---|---|---|
| Subject | TPPs | SE inflation | # TPPs | *Q* | *df* | *p* | Heterogeneity SD | Reliability | Uncorrected | Bonferroni corrected |
| ELA | All TPPs | none | 10 | 19 | 9 | .028 | .040 | .517 | 1 | 0 |
| | | 18% | | 13 | 9 | .146 | .028 | .328 | 0 | 0 |
| | Large TPPs | none | 6 | 12 | 5 | .039 | .050 | .572 | 0 | 0 |
| | | 18% | | 8 | 5 | .136 | .042 | .404 | 0 | 0 |
| Math | All TPPs | none | 8 | 22 | 7 | .003 | .048 | .690 | 1 | 1 |
| | | 18% | | 16 | 7 | .028 | .040 | .554 | 1 | 1 |
| | Large TPPs | none | 7 | 21 | 6 | .002 | .053 | .718 | 1 | 1 |
| | | 18% | | 15 | 6 | .018 | .046 | .608 | 1 | 1 |
| Reading | All TPPs | none | 7 | 14 | 6 | .034 | .052 | .559 | 2 | 0 |
| | | 18% | | 10 | 6 | .135 | .044 | .386 | 0 | 0 |
| | Large TPPs | none | 3 | 2 | 2 | .419 | .006 | .000 | 0 | 0 |
| | | 18% | | 1 | 2 | .536 | .000 | .000 | 0 | 0 |
| Science | All TPPs | none | 7 | 9 | 6 | .160 | .027 | .352 | 1 | 0 |
| | | 18% | | 7 | 6 | .355 | .010 | .098 | 0 | 0 |
| | Large TPPs | none | 2 | 0 | 1 | .606 | .000 | .000 | 0 | 0 |
| | | 18% | | 0 | 1 | .662 | .000 | .000 | 0 | 0 |
| Social Studies | All TPPs | none | 7 | 5 | 6 | .567 | .000 | .000 | 0 | 0 |
| | | 18% | | 3 | 6 | .749 | .000 | .000 | 0 | 0 |
| | Large TPPs | none | 4 | 4 | 3 | .307 | .000 | .169 | 0 | 0 |
| | | 18% | | 3 | 3 | .459 | .000 | .000 | 0 | 0 |