Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

How much does teacher quality vary across teacher preparation programs? Reanalyses from six states

Paul T. von Hippel[a,*], Laura Bellows[b]

[a] LBJ School of Public Affairs, University of Texas, 2315 Red River, Box Y, Austin, TX 78712, United States
[b] Sanford School of Public Policy, Duke University, Box 90239, Durham NC 27708, United States

ARTICLE INFO

Keywords:
Accountability
Heterogeneity
Teacher training

JEL classifications:
I21
I28
C1

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—.01–0.03 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.

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% 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

[*] Corresponding author.
E-mail address: paulvonhippel.utaustin@gmail.com (P.T. von Hippel).

https://doi.org/10.1016/j.econedurev.2018.01.005
Received 20 June 2017; Received in revised form 14 January 2018; Accepted 15 January 2018

between workers selected and trained by different institutions. Among PhD economists, only 10% 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%–9% 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 0.01–0.03 SDs in student test scores.

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 0.01–0.03 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 0.05 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% chance ($1–(1–0.05)^{10}$) 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 larger differences 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.

[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%–10%, 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.

[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.