Economics of Education Review 53 (2016) 31–45

Contents lists available at ScienceDirect

Economics of Education Review

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

Teacher quality differences between teacher preparation programs: How big? How reliable? Which programs are different?

Paul T. von Hippel[a,*], Laura Bellows[b], Cynthia Osborne[a], Jane Arnold Lincove[c], Nick Mills[d]

[a] LBJ School of Public Affairs University of Texas, Austin 2315 Red River, Box Y Austin, TX 78712, United States
[b] Duke University, United States
[c] Tulane University, United States
[d] Formerly LBJ School of Public Affairs University of Texas, Austin 2315 Red River, Box Y Austin, TX 78712, United States

ARTICLE INFO

Article history:
Received 13 March 2015
Revised 12 May 2016
Accepted 12 May 2016
Available online 19 May 2016

JEL classification codes:
I12 (Analysis of Education)
C12 (Hypothesis Testing: General)
C13 (Estimation: General)

Keywords:
Accountability
Teacher training
Heterogeneity

ABSTRACT

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

1. Introduction

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

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

To assess TPP quality, the new accountability systems "focus on student achievement as the primary measure of

* Corresponding author. Tel.: +1 512 537 8112.
E-mail address: paulvonhippel.utaustin@gmail.com (P.T. von Hippel).

http://dx.doi.org/10.1016/j.econedurev.2016.05.002
0272-7757/

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

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

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

A third condition for effective TPP accountability is that we must be able to identify with confidence the individual TPPs that are better or worse than average. Singling out good and bad TPPs is not a trivial matter. It is possible to accept the global hypothesis that TPPs differ in their effects, and yet remain uncertain about which individual TPPs are better or worse. Noise in the estimated TPP differences is just one problem. Another problem is *multiple tests* (Hsu, 1996). We can test each TPP estimate for significance, but if we conduct multiple hypothesis tests at a significance level of .05, then purely by chance we would expect to conclude that 5 of the nearly 100 TPPs in Texas differ significantly from the average—even if all were truly identical. To avoid basing policy decisions on random chance, it is necessary to correct for multiple tests. This correction will inevitably reduce the number of TPPs that appear to be different from average.

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

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

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

While it is possible that the differences between TPPs are larger in some states than in others, it is also possible that the divergent conclusions of past TPP evaluations were due in part to methodological decisions. Past research has highlighted the sensitivity of TPP estimates to decisions about which covariates to include, whether to include school fixed effects (FEs), and how to cluster standard errors (SEs) (Koedel et al., 2015; Lincove, Osborne, Dillon, & Mills, 2014; Mihaly, McCaffrey, Sass, & Lockwood, 2013). There are further modeling issues, such as whether to include random effects (REs) at the teacher or school level (e.g., Gansle et al., 2012a). Once a model has been fit, the methodological decisions are not over. There are a variety of methods that can be used to assess how much noise is present in the estimates, adjust for it, and address the issue of multiple tests.

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

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

In every plausible analysis, we find that the teacher quality differences between TPPs are small, and estimates of those differences consist mostly of noise, even in large TPPs. We also find that few if any TPPs can be confidently flagged as different from average after adjustments are made for multiple tests. These results suggest that TPP accountability systems have limited potential to improve