American Educational Research Journal
February 2023, Vol. 60, No. 1, pp. 183–216
DOI: 10.3102/00028312221137798

Teacher Preparation Programs and Graduates' Growth in Instructional Effectiveness

Emanuele Bardelli
Brown University
Matthew Ronfeldt
University of Michigan
John P. Papay
Brown University

Many prior studies have explored average differences in initial levels of teaching effectiveness among graduates from different teacher preparation programs (TPPs) and the features of preparation that predict these differences. We focus on another important dimension of effectiveness—how graduates from different TPPs improve over time. Examining all graduates from Tennessee TPPs from 2010 to 2018, we find meaningful differences between TPPs in both initial level and early-career growth in teaching effectiveness. We also find that different TPP features explain part of these differences. Yet the features that correlate with initial teaching effectiveness are not the same features that correlate with growth. This article informs policy decisions around TPP evaluation and identifies new directions for future research in TPP effectiveness.

Emanuele Bardelli is a postdoctoral research associate at the National Student Success Accelerator and the Annenberg Institute at Brown University, 164 Angell St., 2nd floor, Providence, RI 02906, USA; email: *emanuele_bardelli@brown.edu*. His research centers on teacher learning, teacher education and teacher induction policy, and quantitative methods to advance education equity.

Matthew Ronfeldt is an associate professor of educational studies at the School of Education, University of Michigan, 610 E. University Ave. Ann Arbor, Michigan 48109, USA; email: *ronfeldt@umich.edu*. His research seeks to understand how to improve teaching quality, particularly in schools and districts that serve marginalized and minoritized students. His work sits at the intersection of educational practice and policy and focuses on teacher preparation, teacher retention, teacher induction, and the assessment of teachers and preparation programs.

John P. Papay is an associate professor of education and economics at Brown University, 164 Angell St., 2nd floor, Providence, RI 02906, USA; email: *john_papay@brown.edu*. His research focuses on policies that affect teachers and educational inequality. His recent work has explored how teachers develop over the course of their careers and how policies affect school staffing.

KEYWORDS: teacher preparation, teacher effectiveness, program features

Preservice teacher education is the main entryway into teaching. It is not surprising, then, that researchers and policymakers have developed ways to evaluate teacher preparation programs (TPPs) using graduates' value-added to student test scores (value-added measure [VAM]) and other workforce outcomes. The general theory behind these policies is that TPP graduates' workforce outcomes are a useful proxy for the quality of preparation provided by TPPs, including the quality of coursework and clinical experiences.

Researchers, on the other hand, have found mixed evidence to support the use of graduates' VAMs to identify more and less successful TPPs (Boyd et al., 2009; Constantine et al., 2009; Darling-Hammond et al., 2005; Glazerman et al., 2006; Goldhaber et al., 2013; Henry, Purtell, et al., 2014; Koedel et al., 2015). In their reanalysis of the data used in some of these prior studies, von Hippel and Bellows (2018) cautioned that most of the observed differences among TPP graduates' VAMs could likely reflect statistical artifacts rather than true differences among graduates' teaching effectiveness. Research into TPP effects on graduates' classroom observation ratings is less developed but has shown promising results. Both Ronfeldt and Campbell (2016) and Bastian et al. (2018) found significant variation between TPPs in terms of their graduates' average observation ratings during their early careers.

These analyses have typically focused on average levels of early-career VAMs or observation ratings among graduates from different preparation programs. However, Goldhaber et al. (2013) argue that TPP's impacts on teacher practice are long-lasting, continuing throughout the first decade in the classroom. And TPPs may not only affect teachers' *level* of effectiveness but also the rate at which they *improve* their practice. That is, teacher practice and effectiveness in raising student test scores develop rapidly over the first few years in the classroom (Atteberry et al., 2015; Harris & Sass, 2014; Henry et al., 2012; Papay & Kraft, 2015; Rockoff, 2004), and there is substantial variation in these growth rates across teachers (Kraft & Papay, 2014). TPPs, then, might differ not only in how they prepare teachers to be initially successful on the job but also in how quickly they improve; for instance, some programs might provide graduates with stronger capacities for reflective practice or ongoing learning that could support more rapid development over time. In fact, principals often report preferring new teachers who are "coachable" rather than initially effective because such teachers can more readily adapt to the school's culture (Giersch & Dong, 2018; Harris et al., 2010).

In this article, we use data from TPP graduates in Tennessee to estimate initial effectiveness and growth in effectiveness over time for graduates from different programs. We focus on VAMs and direct assessments of instructional effectiveness using the state's classroom observation system. On both measures, we find significant variation among TPPs in graduates' initial effectiveness and their early-career improvement trajectories. In some cases,

differences in improvement are sufficiently large to remedy lower levels of initial effectiveness. Moreover, initial effectiveness and subsequent growth appear to be weakly correlated, suggesting that these different outcomes may capture unique aspects of teacher preparation.

Finding substantial between-TPP variation in initial effectiveness and subsequent growth led us to explore which preparation features might relate to the observed variation. Several studies have shown that features of preparation, such as student teaching duration, certification area, and program type are associated with graduates' instructional effectiveness. However, it is not clear that the same features would predict teachers' improvement over time. For example, we may see a trade-off in curricular emphasis between a focus on being prepared on Day 1 and the ability to improve over time through reflective practice. Results from this study suggest that those features that predict initial effectiveness are not always the same ones that predict later growth among TPP graduates. These analyses are exploratory in nature and serve the purpose to showcase how the methods we present in this article could inform future work studying the link between TPP features and graduates' instructional effectiveness outcomes.

Overall, this article highlights the limitations of assessing TPPs simply by examining the instructional effectiveness of graduates in the first year of teaching. We document substantial TPP-specific differences in how teachers improve their instructional practice, and their contributions to student achievement, as they gain experience in the classroom. Importantly, the TPPs whose graduates are most effective in their initial placements are not necessarily those whose graduates improve the most during the next several years in the classroom. This finding has clear implications for TPP evaluation. It suggests that the current TPP evaluation systems might miss an important component of teacher preparation, as average—or even initial—effectiveness does not capture the full range of contributions that TPPs make to their graduates' ultimate influences on student learning and development.

Furthermore, we show that the program features that correlate with graduates' initial effectiveness are not necessarily those that correlate with growth over time. These findings are purely descriptive but suggest that in addition to preparing preservice teachers to enter the classroom as successful novices, TPPs may be able to engage in different practices to set graduates up to take advantage of the substantial on-the-job learning opportunities available in schools. As such, these findings provide potentially important new avenues for future research that can more precisely identify the types of practices that prepare TPP graduates to be successful as novices as well as those that help graduates develop over time.

Literature Review

Differences in the Effectiveness of Graduates From Different TPPs

Individual TPPs differ in many ways, from the students they serve to their unique programmatic design features and characteristics—curriculum, course sequence and structures, clinical placements—intending to support candidate learning and effectiveness. The rise of large, administrative datasets over the past 15 years has allowed researchers to begin to examine quantitatively whether these differences among TPPs relate to differences in outcomes for their graduates. As we review next, this prior research has generally found that TPPs differ in their effects, although, given all the factors that contribute to a teachers' eventual effectiveness, such differences are meaningful but relatively small. Recent studies have also yielded important methodological considerations.

TPP Effects on Student Test Scores

Several studies have found substantial differences in the average instructional effectiveness, as measured by contributions to student test score gains, of graduates from different programs. Boyd et al. (2009) reported that, in New York City schools, the "difference between the average of the institutions and the institution with highest VAMs is approximately 0.07 standard deviations in both math and ELA" (p. 428). Expanding on this work, Goldhaber et al. (2013) found more limited variation for graduates from different TPPs in Washington state. They reported that TPPs explained about "0.01 [standard deviation points] in math and 0.02 in reading" or about "5–12.5% of the standard deviation of the teacher effects" (p. 36). Moreover, they suggested that while TPP effects decay over time, the *half-life* of the average TPP effect is between 11 and 15 years. This suggests that TPPs have long-lasting impacts on their graduates' effectiveness, but that this relationship might wane over time.

TPP Effects on Teachers' Classroom Practice

Efforts to measure teachers' contributions to student learning via test scores have become widespread, but they remain controversial for several important reasons. In particular, scholars have debated whether the statistical models may not fully account for the sorting of students to teachers (e.g., Bacher-Hicks et al., 2017; Bitler et al., 2021; Chetty et al., 2014; Rothstein, 2010). While aggregating VAMs across large groups of teachers, as we do to study TPP effects, eliminates some of these methodological concerns, VAMs do not assess other features of teaching that are important but do not show up in contributions to student test scores (Baker et al., 2010). As a result, scholars have recently begun to examine more direct measures of teachers' instructional practice using classroom observation ratings.

Studies have found robust differences between TPPs using such observation ratings. Using data from Tennessee, Ronfeldt and Campbell (2016) found significant variation in graduates' observation ratings across TPPs, reporting that "institutions [universities, colleges] explained about 2% of variation in graduates' [teacher VAMs] scores and 4% of variation in [observation ratings]; programs [within institutions] explained about 4% of variation on both outcomes" (p. 610). Bastian et al. (2018) found similar results in North Carolina, concluding that "TPPs are significantly associated with the evaluation ratings of their graduates" (p. 442).

Methodological Considerations in Estimating TPP Effects

This literature has provided helpful evidence about the variation across TPPs. However, a series of recent studies have raised important methodological questions and proposed new approaches that we take up in this article. First, several studies have suggested that VAMs may overstate true differences between TPPs in terms of graduates' average instructional effectiveness. The most important critique is that TPP effects may be unable to sufficiently distinguish signal from noise in their estimation of program effects on graduates' instructional effectiveness. Koedel et al. (2015) argued that past work has not taken into account teacher sampling and reported that most of the variation in teacher effectiveness is explained by differences within programs rather than differences between programs.

Von Hippel et al. (2016) made these challenges more explicit, using meta-analytic methods to show that most of the variation in graduates' teacher effectiveness outcomes for Texas TPPs could be explained by random measurement error. They showed that TPP effects for Texas programs follow the expected natural distribution under the assumption that there are no TPP effects and that the observed differences are due to naturally occurring variation. In a follow-up paper, von Hippel and Bellows (2018) reanalyzed data from NYC, Washington State, Missouri, and Florida, and again found little evidence of differences across TPPs beyond expected random noise. These studies suggest caution in concluding that observed variation in workforce outcomes for TPP graduates is associated with differences across TPPs, unless careful consideration is given to both the methodological approach taken to estimate these effects and to the extent to which natural variation in TPP effects could explain any observed difference in graduates' outcomes.

Mihaly et al. (2013) raised an additional concern with implications for how to address the clustering of program graduates in certain schools once they enter the workforce. They highlighted the challenge of accounting for the schools in which TPP graduates end up teaching, arguing that traditional school fixed effects modeling approaches could lead to variance inflation as "school fixed effects can be collinear with the program effects in the model when graduates of some programs never teach with graduates of other

programs and groups of programs have many connections within the groups but few outside the group" (p. 486). They found that including school fixed effects led to losing up to 63% of the within-TPP variability in teacher effectiveness, leading to the TPP variance in models with fixed effects to be almost twice the variance in models without TPP fixed effects.

TPP Effects on Teachers' Effectiveness at Entry Versus Over Time

The vast majority of studies linking preparation features to graduates' instructional effectiveness have focused on *levels* of effectiveness for graduates in their first few years of in-service teaching. The assumption is that "effective" programs are ones whose graduates are more instructionally effective right away. However, several scholars have argued for teacher education that prepares teachers for not just immediate effectiveness but the ability to *improve* practice over time. For example, scholars argue that TPPs should prepare teachers for deliberation on and critical examination of actions (Kennedy, 1987), learning in and from practice through reflection and inquiry (Cochran-Smith & Lytle, 1993; Schön, 1983). In fact, Dewey (1904) argued that an emphasis on immediate skill can actually stunt growth, arguing instead for preparing graduates to become "students of teaching." If programs are successfully preparing prospective teachers for growth rather than immediate skills, then research focusing only on initial effectiveness will miss their impacts.

Several studies have addressed this question implicitly, focusing on how early career teachers from alternative certification programs compare to those from traditional teacher education programs, in aggregate. For example, Kane et al. (2008) find that New York City Teaching Fellows graduates appear to become more effective than their peers from traditional programs over the first few years of their career, while Papay et al. (2012) find similar results for graduates of the Boston Teacher Residency program. We know of no studies that explore how the relative effectiveness of graduates from different traditional TPPs evolves over time.

Features of Preparation

While many researchers focus on the aggregate effects of TPPs on graduate outcomes, during the past decade a growing number of studies have examined how program features that vary within and across programs predict graduate workforce outcomes (see Ronfeldt, 2021, for a review of this literature). This shift towards focusing on the effects of features rather than whole programs has been driven, at least in part, by the mixed evidence about whether meaningful differences exist between programs in terms of graduates' average instructional effectiveness.

These prior analyses have focused almost exclusively on whether preparation features predict average teaching effectiveness but not growth over time. We model the relationships of some program features with both initial

levels of instructional effectiveness as well as growth over time, and test whether the features that predict initial effectiveness also predict later growth. Our contributions here are meant to be largely methodological and exploratory in nature; we are not aiming to make substantive contributions about which kinds of preparation matter most. Instead, we select four features (a) that represent various program dimensions, (b) that are shown in prior literature to be related to graduates' instructional effectiveness, and (c) for which we have information for all programs in our sample.

Next, we describe the four features we selected and summarize what we have learned from prior literature linking these features to graduates' instructional effectiveness.

Program Type (Endorsement Area; Graduate/Undergraduate)

Prior literature has suggested that graduates from different kinds of program types, including endorsement area and whether a program is graduate or undergraduate, have different average instructional effectiveness. It is important to acknowledge here that these differences in graduates' instructional effectiveness between program types likely stem from two different sources—differences in candidate recruitment/selection and differences in preparation experiences. If we find, as a hypothetical, that graduates from secondary math programs outperform peers from elementary programs, then it could be because the former programs recruit/admit more promising candidates or that these programs offer different and better kinds of preparation (e.g., stronger clinical mentors; better coursework), or both. It is not our goal to differentiate these sources. Instead, following Boyd et al. (2009), we conceptualize both recruitment and preparation to be part of the design and purpose of teacher preparation.

Bardelli and Ronfeldt (2020) examine workforce outcomes among early-career Tennessee teachers who received certification in different high-needs endorsement areas (HNEAs) which included STEM (science, technology, engineering, and mathematics), special education (SpEd), and bilingual/English as a second language (ESL) endorsements. The authors found that SpEd-endorsed teachers tended to have lower VAMs than HNEA teachers, while STEM and bilingual/ESL teachers generally had comparable VAMs; the only case where HNEA outperformed non-HNEA teachers was that teachers with STEM endorsements had greater mathematics VAMs. This finding is consistent with Ronfeldt (2015), who found that teachers with mathematics certification tend to have stronger achievement gains in mathematics than teachers who are certified in other areas. However, when Bardelli and Ronfeldt (2020) considered observation ratings rather than VAMs, they found the opposite to be true—STEM (and SpEd) teachers received significantly lower observation ratings than non-HNEA peers. Henry et al. (2012) is the only prior study, to our knowledge, that has considered differences between

certification area in terms of teachers' growth in, rather than level of, instructional effectiveness. Specifically, they find that VAMs of science teachers grow at about double the rate of VAMs of mathematics teachers and about four times the rate of VAMs of non-STEM teachers.

A number of studies have compared average teaching effectiveness of graduates from undergraduate versus graduate/postbaccalaureate programs, showing mixed findings. Henry, Bastian, et al. (2014) found that undergraduate programs outperformed graduate programs in terms of middle grades mathematics value-added. In contrast, Henry, Purtell, et al. (2014) found that graduates from in-state public graduate programs had higher VAMs in high school mathematics than did graduates from in-state public undergraduate programs. In Washington, Cowan et al. (2017) found no differences between undergraduate and graduate programs regarding mathematics VAMs. Two studies found graduate programs to outperform undergraduate programs in terms of ELA VAMs across grade levels in Washington (Cowan et al., 2017) and in middle grades in North Carolina (Henry, Bastian, et al., 2014). However, the third study found no differences (Henry, Purtell, et al., 2014).

Both Henry, Purtell, et al. (2014) and Henry, Bastian, et al. (2014) explored TPP effects on test scores in other subjects in North Carolina. The former found in-state private graduate program completers to have better high school science VAMs than in-state public undergraduate programs. Differences between graduate and undergraduate programs were statistically similar in the latter study. Both studies found no differences between graduate and undergraduate program effectiveness in other subject areas.

Clinical Experiences

There is substantial evidence that the features of clinical experiences matter (Ronfeldt, 2021). Given that we only consider a small set of preparation features, we focus on two main features of clinical experiences—student teaching duration and student teaching placement type. We hope future studies will consider a more comprehensive set of features, including those related to coursework.

Student Teaching Duration

While a number of studies have found that graduates who complete longer student teaching experiences report feeling better prepared to teach (California State University, 2002; Ronfeldt et al., 2014, 2020; Ronfeldt & Reininger, 2012), there is little evidence that completing more student teaching is related to instructional effectiveness. Three studies have explored the relationship between measures of student teaching duration and graduates' VAMs, finding inconsistent, null, or mixed results (Boyd et al., 2009; Preston, 2017; Ronfeldt, 2015). Ronfeldt and colleagues (2020) found no relationship between student teaching duration and graduates' observation ratings.

Clinical Placement Type

The U.S. Department of Education classifies clinical experiences in three categories: (a) student teaching, (b) internship, and (c) job-embedded (see Title II, Section 202 (d)(2) of the Higher Education Opportunity Act [P. L. 110-315] for the specific guidelines). *Student teaching* placements reflect more traditional, semester-long placements that typically occur as a capstone experience at the end of preparation. Similar to those used in residency programs, *internship* placements typically include a full year in the same classroom co-teaching with the same cooperating/mentor teacher. Both *student teaching* and *internship* placements serve preservice candidates prior to becoming an in-service teacher of record. By contrast, *job-embedded* placements, which are commonly associated with alternative route pathways, are completed as an in-service teacher of record, where teachers complete clinical requirements while being legally responsible for students. Job-embedded placements typically require individuals to have some form of provisional certification, as the completion of the job-embedded placement is required as part of formal certification. All three placements typically include a cooperating/mentor teacher. For student teaching and internship placements the mentor teacher is the primary teacher of record who actively supports the learning of P–12 students, though the degree to which they play a lead or supporting role can vary. By contrast, in job-embedded placements, the learning teachers' mentors are not typically teaching alongside them in their classrooms, though they may observe and give feedback from time to time.

We are not aware of any literature that has examined the degree to which these clinical placement types, on their own, are related to initial teaching effectiveness or later growth. That said, there are bodies of literature that are related. First, the literature on the duration of placements (see prior section) has some relevance given that traditional student teaching tends to be a semester while internships are typically a full year; job-embedded placements are typically 2 years in duration. However, there are other differences beyond duration that obscure this distinction. For example, job-embedded placements are not just longer but are completed as teacher of record, which likely affords and constrains different learning (to teach) opportunities; moreover, job-embedded placements typically occur following shorter-term, preservice field experiences that often occur over the preceding summer.

The other, related body of literature has compared routes or pathways of entry and average differences in teaching effectiveness, typically measured in terms of VAMs. There is a long history of studies comparing the VAMs of graduates from alternative versus traditional route programs, finding mixed results (Grossman & Loeb, 2008). Likewise, there is an emerging body of literature comparing graduates of residency and nonresidency programs (Papay et al., 2012). Literature comparing pathways/routes of entry, though, are unable to disentangle effects of placement type from other features of preparation

that differ between pathways/routes, such as the timing of coursework and the presence and role of the mentor/cooperating teacher.

Interpretation: Correlation, Selection, and Threats to Causal Inference

Interpreting the relationships we observe between a given program feature and graduates' instructional effectiveness requires the consideration of many forms of selection (Goldhaber & Ronfeldt, 2020). First, different kinds of candidates select into programs that have different kinds of features. For example, if more promising candidates tend to select into programs with longer student teaching experiences, we cannot disentangle whether any positive observed relationship between student teaching duration and instructional effectiveness is due to longer student teaching or candidate characteristics. Second, certain program features are likely correlated with other features. Continuing the example above, if programs that require longer student teaching duration also tend to recruit more instructionally effective mentor teachers, we would again be unable to know if the observed relationship is explained by duration or cooperating teacher effectiveness. Third, TPP effects could be due to the kind of schools and districts where their graduates eventually work. Evidence suggests that teacher labor markets are quite local (Boyd et al., 2005), that student teaching placements can be an important pathway to jobs (Bartanen & Kwok, 2020; Cannata, 2011; Jabbar et al., 2020), and that teacher candidates prepared in more selective TPPs are more likely to be hired than their peers from less selective TPPs (Jacob et al., 2018; Reininger, 2012; Rockoff et al., 2011). Thus, TPP-level estimates may conflate the effects of teacher preparation with unrelated local school and district policies and practices. Finally, some TPPs (e.g., Teach for America) offer training and support to their graduates during their first few years of teaching experience. This continued support might itself support ongoing teacher development and interact with school-based induction practices for early career teachers, leading to better teacher effectiveness over time.

Given these various forms of selection, it is important to underscore that the findings in this article are descriptive and correlational in nature; they cannot be interpreted as causal. Even so, establishing correlational evidence is often a promising foundational step for the design and development of future causal analyses.

Research Questions

The research questions that guide this article are

RQ 1: How much of the variance in early career teachers' initial effectiveness and growth do teacher preparation programs explain?

RQ 2: Are there differences between programs in terms of graduates' initial effectiveness and growth during their early teaching careers?

RQ 3: Which program-level features predict graduates' initial effectiveness and growth during their early teaching careers? Do the same features that predict initial effectiveness also predict growth?

Methods

Data and Measures

Data for this project comes from Tennessee Statewide Longitudinal Data System. This system includes comprehensive data on student enrollment, annual student test scores, teacher employment, school characteristics, and teacher effectiveness. We draw information about TPP features—program level, clinical placement type, and student teaching type—from the U.S. Department of Education Title II dataset.

Sample

We focus on the universe of 43,917 new teachers in Tennessee from 2010 to 2018. Our sample includes TPP graduates from in-state programs that we can link to teacher preparation data and teachers in their first year of teaching who we cannot link to teacher preparation data. The latter group includes teachers prepared by out-of-state TPPs and teachers who are working towards their teaching credential without having completed a TPP yet. We report summary statistics for our analytic sample in Table 1. Early career teachers who we can link to TPP records appear to be similar to early-career teachers who cannot link in both instructional effectiveness[1] and demographic characteristics. While we do not focus on the latter group, we include them in our analytic sample to help estimate the underlying relationships between our covariates and outcomes.

Focal Outcomes

We focus on two outcomes to assess teacher effectiveness: (a) observation ratings of teachers and (b) VAMs of teachers' contributions to student test scores. These measures are collected as part of Tennessee's teacher evaluation system and are used to calculate a summative evaluation score for each teacher in the state.

Observation ratings are based on the Tennessee Educator Acceleration Model (TEAM) observation rubric. This rubric organizes 24 indicators into four domains: instruction, planning, environment, and professionalism. About 80% of all teachers in the state are evaluated using TEAM each year by a trained evaluator, typically a school administrator. These evaluations are usually based on two or more classroom visits and debrief sessions with the evaluator. We average indicator-level scores for each teacher in each school year to get a summative rating of the teacher's classroom practice in

Table 1
Descriptive Statistics

	Teacher Level			Teacher-Year Level		
	All	With TPP Data	Without TPP Data	All	With TPP Data	Without TPP Data
Outcomes of interest						
Observation ratings	–0.401	–0.344	–0.444	–0.249	–0.187	–0.286
	(0.923)	(0.959)	(0.891)	(0.985)	(0.998)	(0.976)
VAMs	–0.125	–0.146	–0.112	–0.027	–0.052	–0.016
	(0.869)	(0.900)	(0.850)	(1.014)	(1.003)	(1.018)
Teacher covariates						
Years of experience	1.899	2.153	1.702	2.485	2.791	2.305
	(1.469)	(1.708)	(1.216)	(2.096)	(2.166)	(2.032)
Initial years of experience	0.919	0.942	0.898	0.919	0.923	0.917
	(1.777)	(1.348)	(2.093)	(1.728)	(1.168)	(2.000)
Female	0.767	0.766	0.767	0.770	0.779	0.765
	(0.423)	(0.423)	(0.423)	(0.421)	(0.415)	(0.424)
Asian or Pacific Islander	0.008	0.005	0.010	0.007	0.004	0.009
	(0.089)	(0.072)	(0.100)	(0.085)	(0.067)	(0.094)
Black or African American	0.105	0.108	0.103	0.102	0.095	0.106
	(0.306)	(0.311)	(0.302)	(0.303)	(0.294)	(0.308)
Hispanic or Latinx	0.009	0.006	0.012	0.009	0.005	0.012
	(0.095)	(0.076)	(0.107)	(0.095)	(0.067)	(0.108)
White	0.794	0.729	0.845	0.830	0.798	0.848
	(0.401)	(0.444)	(0.355)	(0.376)	(0.402)	(0.359)
Other	0.009	0.000	0.015	0.007	0.000	0.012
	(0.091)	(0.000)	(0.121)	(0.086)	(0.000)	(0.108)
Average number of obs.	4.040	3.420	4.521	—	—	—
	(2.414)	(2.143)	(2.501)			
N	43,917	19,171	24,746	177,429	65,556	111,873

Note. This table reports descriptive statistics for the teachers in our sample. The outcomes of interests are standardized using the statewide sample of all teachers and within school year. Teacher level statistics report averages within a teacher. Teacher-year estimates allow for multiple observations for each teacher and for time-varying variables to take different values within the same teacher. Standard deviations are in parentheses.

a given year. The state began its evaluation system during the 2011–2012 school year. For teachers who are not evaluated using the TEAM rubric, we rely on the Tennessee Department of Education's TEAM-equated scores.

The state's teacher accountability system also includes individual value-added scores for the subset of teachers who teach tested grades and subjects—math, English language arts (ELA), science, or social studies in Grades 3 through 8 and most math and ELA teachers in high school. This includes roughly 44% percent of the teachers who work in the state. The state calculates these scores following the Education Value-Added Assessment System (EVAAS) model (Vosters et al., 2018) that seeks to isolate each teacher's contribution to students' test scores. We create average composite VAMs for teachers who have scores in multiple tested subjects. VAMs are available starting from the 2010–2011 school year.

We standardize the observation ratings and VAMs within each school year using the statewide sample of all teachers to have mean 0 and standard deviation of 1. This ensures that year-to-year shocks in outcomes are taken out from the teacher-level scores and allows us to interpret a score of 0 as the score of an average teacher in the state.

Teaching Experience

As we focus on modeling growth in effectiveness over time, years of teaching experience is the identifying covariate in our models. We define experience as the number of years as a classroom teacher of record since entering the TPP program. For program completers (PCs) with traditional or internship clinical placement types (more on this below), the first year of experience occurs after graduating from a TPP. For job-embedded clinical placement types, however, PCs are serving as teachers of record prior to graduating; thus, their first year of experience is concurrent with the year in which they complete their TPP. Because some PCs have classroom teaching experience before entering their TPPs, we separately code and include in our models the years of prior teaching experience a PC is reported having in the state's personnel data before entering the TPP.

Program Features

We combine data from the state's program completer dataset with U.S. Department of Education Title II data to compile a dataset of program features for all TPPs in our dataset.

The program completer dataset includes information on the endorsement areas that graduates receive after completing their program, the types of programs they complete, and clinical placement that they experienced. We recode individual endorsements into five categories: elementary education, secondary STEM education, secondary non-STEM education, special

education, and other certification areas. The last group includes endorsements that span all K–12 grades such as music or physical education.

We code program type using the degree level that PCs are reported receiving with their teacher certification. These include baccalaureate programs, post-baccalaureate programs, and nondegree teacher certificate programs.

The state follows U.S. Department of Education categorization of clinical placements as follows: (a) traditional student teaching as a field placement up to a semester in length, (b) internship as a year-long field placement, and (c) job-embedded as a field placement that is contingent on employment as a teacher of record and concurrent to teacher preparation coursework. Most of the TPPs in the state offer job-embedded pathways alongside traditional student teaching placements, as Tennessee has had these programs in place since the early 2000s.[2] Overall, we observe about 30% of all PCs in the state completing a job-embedded placement. Fewer programs offer internship placements—about 5% of PCs experience this clinical placement type—and these placements are usually associated with residency programs offered alongside traditional clinical placements within the same institution. The remaining 65% of teachers complete a traditional student teaching placement.

Finally, we use program-level Title II data to measure TPPs' student teaching length. TPP leaders self-report these data yearly to the U.S. Department of Education, which then makes them publicly available online. We divide this variable into terciles and compare programs in the upper and middle tercile to programs in the bottom third of the distribution. We also separate programs that we cannot link to these Title II data in a separate category to ensure that the estimation sample remains consistent across all by-feature analyses.

Design and Analysis

Our key questions involve exploring how teachers improve their effectiveness over time. Our central analyses are purely descriptive. We seek to understand whether teachers who attend certain TPPs are initially more effective and improve at greater rates early in their careers than teachers who attend other TPPs. Ideally, we would want to draw causal conclusions about the impact of TPP practices on these outcomes. However, we cannot make these causal links because recruitment, TPP experiences, and TPP graduate placement are all confounded. Therefore, we caution the reader not to interpret our results causally. That is, our analyses describe the association between attending a given TPP and later effectiveness.

Analysis

RQ 1: Variation in Instructional Effectiveness Explained by TPPs

We are interested in exploring whether the TPPs that graduates attended explain a significant part of the variation in teachers' initial effectiveness and

improvement early in the career. We use longitudinal, multilevel models to decompose the variance in our outcomes of interest by nesting teacher-year observations within teachers, and teachers within TPPs. Conceptually, this modeling approach allows us to separate the variance in teaching effectiveness measures into three parts: a part related to teacher-level factors, a part related to TPP-level factors, and a residual related to time-varying factors within teacher.

More formally, we estimate the following model:

$$Var(Y_{itp}) = \tau_i + \tau_p + \sigma^2,$$

for teacher i in year t in program p. Here, τ_{i} and τ_p are the variance components explained at the teacher and program level respectively, while σ^2 is the residual variance.

We calculate the ICC values as the proportion of variance that is explained by intraclass correlations as

$$ICC_p = \frac{\tau_p}{\tau_{\mathrm{i}} + \tau_p + \sigma^2}.$$

This estimates the fraction of variance that is explained at the TPP level. We interpret this as the upper bound estimates for the portion of variance explained by TPPs that we can observe in the rest of our analyses.

RQ 2: TPP Relationships With Initial Instructional Effectiveness and Growth Rates

We use a longitudinal, multilevel model to estimate the association between attending a given TPP and instructional effectiveness, initially and over time. We note here that our preferred models for the next two research questions use a different random effects structure than the models for RQ1 for two reasons. First, we include TPPs as fixed effects in these new models, which would make the inclusion of a TPP-level random effect redundant. Second, we include school and district random effects to model the nesting of teachers within schools and districts.

Our preferred model is the following multilevel model:

$$Y_{isdt} = \beta_0 + \beta' \cdot f(exp)_{isdt} + \gamma' \cdot \ TPP_i + \delta' \cdot f(exp)_{isdt} \times TPP_i + (\nu_d + \nu_{sd} + \nu_{isd}) + \epsilon_{isdt},$$

where Y_{isdt} is either the standardized within each year observation score or VAMs for teacher i, in school s and district d, during year t. $f(exp)_{isdt}$ is a function of years of experience for teacher i, described below. TPP_i is a set of indicators for each TPP in the state. These variables take the value of 1 if teacher i has graduated from a given program and 0 otherwise. We finally interact $f(exp)_{isdt} \times TPP_i$ to allow for the TPP coefficients to vary across experience levels.

γ and δ are vectors of coefficients of interest. These coefficients are TPP fixed intercepts and slopes and estimate the association between attending a TPP and instructional effectiveness outcomes as the estimates vary across experience levels. γ captures differences among first-year effectiveness for graduates from different TPPs. δ captures the differential growth rates for TPP graduates in later-experience years. It is worth to note here that these regression coefficients report the deviation from the state average growth rate estimated by β. Therefore, we report the coefficient $\gamma + \delta$ for each TPP and spline section in the result section to help with the interpretation of our results.

We partition error variance into four terms using a three-level multilevel model. ν_d, ν_{sd}, and ν_{isd} are nested random intercepts terms for each school district d, school s, and teacher i, respectively. These terms capture the nested nature of the data and allow for the initial teacher effectiveness level to vary for each individual teacher, school, or district. Intuitively, these random intercept terms capture any unobserved contribution to effectiveness that is are not correlated with TPPs or the other covariates included in the models. These factors could include individual teacher beliefs and dispositions towards instruction, school hiring preferences, or district induction practices. Finally, ϵ_{isdt} is the year-specific idiosyncratic error term.

We model the teacher experience profile in two ways. First, we use a piecewise linear spline to calculate the average difference in effectiveness for given parts of the teaching career. This spline has three parts: 0 to 2, 3 to 5, and 6 to 9 years of experience. Using a spline allows us to estimate parsimoniously TPP coefficients as they differ across experience levels. Our spline approach assumes that the teacher improvement follows a linear functional form on small parts of the experience profile. However, we also relax this assumption by modeling experience using indicator variables for each year of experience. While this specification does not assume any functional form for the relationship between teacher experience and outcomes, it also suffers from the risk to overfit our model to our data. We present a visual comparison of the results of these two alternative specifications for experience in Appendix Figure 1, available online. Because these two specifications provide virtually identical estimates, we use the linear spline specification for both interpretability and parsimony.

RQ3: TPP Features That Predict Graduates' Initial Instructional Effectiveness and Growth Trajectories

We modify the model that we discussed for RQ 2 to estimate the relationship between program-level features and graduates' initial effectiveness and growth trajectories. We replace the TPP indicator variables with features indicator variables. These feature variables group together programs that share common attributes (i.e., license type, degree level, clinical placement, student teaching length).

Bias Versus Precision in Our Random Effects Models

PCs from a given TPP may cluster in different schools. Failing to account for this sorting can lead to bias, as we may misinterpret any average differences across schools as being associated with the TPP. However, traditional approaches that control for school fixed effects may overcorrect by only comparing teachers within the same school. As a result, we use school random effects models in an attempt to balance bias and precision in our models. Conceptually, random effects models take the fixed effects estimates and shrink them towards the population mean for clusters with high variation.

As a concrete example, assume that school A hires a few PCs every year and that these PCs have higher than average effectiveness. A school fixed effects model only relies on within-school variation, netting out the high average effectiveness for these PCs.

A random effects model shrinks the estimated school effects towards the population grand mean, particularly for schools with extreme fixed effects values or schools with few observations. This makes random effects models better suited in our case, where we could have high-performing graduates from the same TPP working in the same set of schools (e.g., a program that recruits promising candidates that are placed in few selected schools) or TPPs that supply new teachers to the same schools (e.g., TPPs working in rural areas of the state).

Random effect models, however, have important limitations. For example, they are subject to bias when omitted variables are correlated with predictors in the regression model. To partially address this concern, we include a set of school-level covariates in our models, including student body characteristics, such as percentage of students in race/ethnicity identity groups, percentage of students who qualify for reduced-priced/free lunch, percentage of students with disabilities, and percentage of students who are classified as English language learners; as well as a 3-year average for the percentage of teacher turnover, a proxy for school working conditions (Ronfeldt, 2012). The inclusion of covariates could ameliorate, at least in part, concerns with omitted variable bias and address some of the concerns with sorting of TPP graduates into different schools and districts.

As a robustness check, we use a two-stage model to estimate the extent to which schools could explain the variance among TPPs. In the first stage, we residualize teacher evaluation outcomes with regression models that include school and year fixed effects. In the second stage, we use these residualized teacher evaluation scores as the outcome variable for the longitudinal, multilevel models described above. We interpret the results from these two-stage analyses to be the likely lower bound for the portion of variance explained by TPPs.

We generally find that these two-stage models reduce the variance that TPPs explain by about half and that most of the adjustment happens at the

middle of the distribution of TPP effects. In other words, these two-stage models adjust the middle of the TPP distribution towards the state mean while keeping the effects on tails of the distribution mostly intact. As we observe most of the differences in TPP coefficients at the tails of the distribution, we expect that this adjustment is not needed in our case and that district and school random effects models are an appropriate modeling approach for the rest of this paper. Moreover, the results from the two-stage models are qualitatively similar to the ones from our preferred modeling approach (see Appendix Figure 3).

Robustness Checks

One of the main concerns with our modeling approach is that teacher attrition among TPP graduates could explain the differences in growth rates that we observe. The intuition behind this concern is that the average growth rate for a TPP could seem to increase not because its graduates are improving at faster rates but because of changes in the distribution of graduates' instructional effectiveness due to the selective attrition of lower-performing teachers. If this mechanism is at play, we would conflate growth from selective attrition with growth related to attending a particular TPP.

We address these concerns in two ways. First, use a two-step inverse probability weight (IPW) model. In the first step, we estimate the probability of leaving each year, using a probit model that includes the same variables as our preferred models (i.e., experience spline, school covariates, PC covariates, and TPP fixed effects). We then predict the expected probability of leaving and use this to calculate IPW. We use these weights in regression models analogous to the ones that we use in the article, although we estimate ordinary least squares (OLS) regression models where we use the IPWs as importance weights and cluster the standard errors at the teacher level. Results for these analyses show that, at least with the IPW adjustment that we used, differential employment rates do not explain the differences we observe among TPPs. On the contrary, these analyses could suggest that adjusting the workforce outcomes for employment rates would increase the variance observed among TPPs.

Second, we compare the growth in evaluation scores for teachers that stay and teachers that leave in any school year by TPP. This test allows us to assess the extent to which our growth models could be biased by worse (or better) performers' career decisions. Formally, we fit the following regression model:

$$\Delta \text{Eval}_{ip} = \beta_0 + \beta_1 \cdot \text{Leaver}_i + \beta_2 \cdot \text{TPP}_p + \beta_3 \cdot \text{Leaver}_i \times \text{TPP}_p + \text{Controls}_{ip} + \sigma_i + \gamma_i + \epsilon_{ip},$$

where ΔEval_{ip} is the difference between the current year and the previous year evaluation scores, *Leaver* is an indicator variable for whether a teacher leaves in a given school year, TPP_p is a set of indicator variables for each teacher preparation program in the state, $Controls_{ip}$ is the same set of control variables that we include in our preferred models (e.g., years of teaching

Table 2
Variance Decomposition

	Observation Ratings		Value-Added Measures	
	Raw	Two-Step	Raw	Two-Step
TPP	0.034	0.015	0.014	0.004
Teacher << TPP	0.643	0.578	0.381	0.312

Note. Null models do not include any covariates. The covariance structure nests teacher-year observations within teacher (Level 2) and within TPP (Level 3). Residualized outcomes are calculated in two stages. In the first stage, we calculate the residuals of school and year fixed effects models. In the second stage, we use these residuals as the outcome variable for the mixed effects regressions.

experience, school-level characteristics, and graduation cohort), σ_i is a school fixed effect, γ_i is a school year fixed effect, and ϵ_{ip} is the residual term. All standard errors are robust. We do not find evidence that there is significant heterogeneity in the growth rates for stayers and leavers across TPPs. The only way for attrition to bias our results is to observe heterogeneity in the growth rates of stayers or leavers across TPPs; since we do not observe this, it is unlikely that differential attrition among TPPs is a significant source of bias in our estimates.

For these reasons, we have decided to report the unadjusted in the article as they appear to be more conservative than the employment-adjusted estimates.

Results

We report our results in three sections, following our research questions. In RQ 1, we find that TPPs explain about 3% of the variance in early career observation scores for TPP graduates and about 1.5% of the variance in VAMs. We expand on these exploratory findings in RQ 2, showing substantial differences between top- and bottom-quintile TPPs in both initial effectiveness and early-career growth of graduates. Finally, we examine which features of preparation are related to initial effectiveness and later growth; we find that the features which predict initial effectiveness differ from those that predict later growth, though some features explain both.

RQ 1: How Much of the Variance in Early Career Teachers' Initial Effectiveness and Growth Is Explained by Teacher Preparation Programs?

TPPs explain a significant portion of the variance in our outcomes of interest. In Table 2, we report the results of this decomposition from our multilevel models, with observation ratings on the left and VAMs on the right. The

first two rows report the variance components for the two levels in our multilevel models—teachers and TPPs. Differences across TPPs explain 3.4% of the variance in observation ratings and 1.4% of the variance in VAMs or, in other words, about 4% to 5% of the variance explained by teachers.

Because of concerns about the sorting of TPP graduates to specific schools, we also estimate the variance components using an alternative specification for our outcomes of interest. Here, we residualize observation ratings and VAMs from models that control for school fixed effects. This approach reduces the estimated TPP variance by at least half. This could suggest that the graduates from the same TPP tend to be employed in the same or similar schools, which makes it difficult to separate analytically TPP effects from school effects. As we are unable to parse these separate effects with our dataset, we see the results of the residualized outcomes as a possible lower bound for the estimates of the TPP variance components.

RQ 2: Are There Differences Between Programs in Terms of Graduates' Initial Effectiveness and Growth During Their Early Teaching Careers?

We first estimate the average statewide growth trajectories for observation ratings and VAMs. In Table 3, we report the results of two models: (a) null models which estimate the average effectiveness for TPP graduates and provide unconditional estimates of the variance components in our multilevel models, and (b) growth curve models which estimate the first year of instructional effectiveness as well as early career growth slopes. We report these for observation ratings and VAMs. Consistent with the literature, average annual growth is greater early in graduates' careers for both outcomes and tapers off with time (Atteberry et al., 2015; Harris & Sass, 2014; Kraft et al., 2020; Papay & Kraft, 2015; Papay & Laski, 2020; Rockoff, 2004).

Figure 1 displays estimates of TPP effects, illustrating the variation in initial effectiveness by TPP. The range is approximately 0.6 standard deviations for initial observation ratings and 0.3 standard deviations for VAMs. This range is approximately equal to the growth we observe in 2 years of teaching experience for observation ratings and about 3 years of experience for VAMs. Said another way, we observe significant differences in the initial effectiveness of graduates from different TPPs, and these differences could be equal to several years of teaching experience.

Von Hippel and Bellows (2018) have argued that estimates of the effects of TPPs on VAMs could be just due to statistical noise. They propose a test to measure the extent to which TPP effects follow the expected null distribution given random measurement error. We report the results of these tests in Appendix Figure 2, available online, where each panel reports on either initial effectiveness or growth for one of our outcomes. We represent the point estimate for each TPP as a black diamond; display 95% confidence intervals as the solid whisker around the point estimate; and, following von Hippel and

Table 3
Growth Estimates for Observation Ratings and Value-Added Measures

	Observation Ratings		Value-Added Measures	
	(1) Null Model	(2) Growth Curve	(3) Null Model	(4) Growth Curve
Average effectiveness	–0.295***		–0.121***	
	(0.028)		(0.017)	
First-year average		–0.794***		–0.380***
		(0.048)		(0.114)
Growth during Years 0–2		0.256***		0.123***
		(0.003)		(0.008)
Growth during Years 3–5		0.079***		0.009*
		(0.002)		(0.005)
Growth during Years 6–9		0.018***		–0.013*
		(0.003)		(0.008)
School covariates	No	Yes	No	Yes
N	123,553	123,553	47,491	47,487
Covariance structure				
District - Intercept	0.085	0.079	0.018	0.011
	(0.013)	(0.012)	(0.004)	(0.004)
School - Intercept	0.151	0.134	0.090	0.087
	(0.007)	(0.006)	(0.006)	(0.006)
Teacher - Intercept	0.495	0.440	0.343	0.333
	(0.005)	(0.004)	(0.007)	(0.007)
Residual	0.322	0.293	0.598	0.593
	(0.002)	(0.002)	(0.005)	(0.005)

Note. Standard errors in parentheses. School covariates include student body characteristics, such as percentage of students in race/ethnicity identity groups, percentage of students who qualify for reduced-priced/free lunch, percentage of students with disabilities, and percentage of students who are classified as English language learners; as well as a 3-year average for the percentage of teacher turnover.
*$p < .1$. ***$p < .001$.

Bellows (2018), report the Bonferroni adjusted 95% confidence intervals as dashed whisker and the gray dots. Below each plot, we report the Cochran's Q statistic, its probability value, an estimate for the heterogeneity variance (τ) in the outcome that TTPs could explain, and an estimate of reliability (ρ) or the fraction of variance explained by systematic differences across TPPs instead of error.

For all tests, we reject the null hypothesis that the distribution of our estimates for TPP initial effectiveness follows the null distribution, suggesting statistically significant differences across TPPs in both initial effectiveness and early career growth rates. Our estimates for initial observation ratings seem

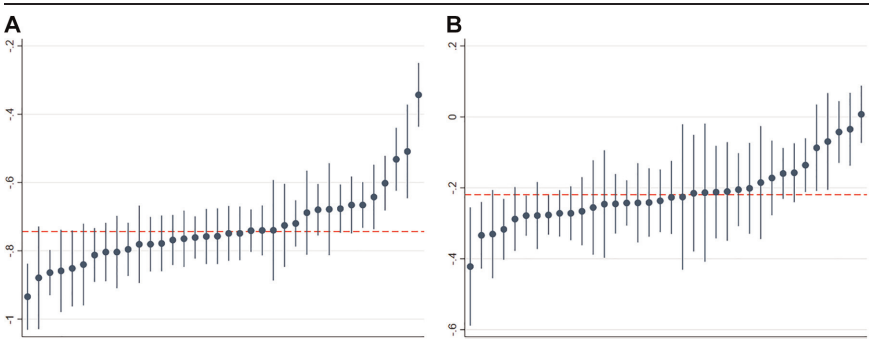

Figure 1. **Differences in Initial Performance by Teacher Preparation Program: (A) Observation Ratings and (B) Teacher Value-Added Measures.**

to be somewhat more reliable in distinguishing TPPs, $\rho = 0.79$, than our estimates for VAMs, $\rho = 0.70$.

To illustrate the implications of these findings, we display the growth trajectories for graduates from four selected TPPs in Appendix Figure 3. We highlight two patterns. First, TPPs whose graduates start at similar levels of effectiveness have different growth trajectories. For example, graduates of the yellow and red TPPs start at similar levels of initial effectiveness. However, graduates from the yellow TPP improve at faster rates than graduates from the red program.

Do Programs Whose Graduates Have Better Initial Effectiveness Also Have Better Growth Trajectories?

Table 4 reports the correlations between the initial effectiveness level and the growth trajectories for the first two terms of the experience spline (i.e., 0–2 and 3–5 years of experience). Overall, we find mixed evidence that these three variables are correlated. For observation ratings, initial effectiveness is positively correlated with the growth rate during the first 3 years of teaching ($r = 0.28$). This suggests that teachers that receive initially higher observation ratings tend to also improve at greater rates early in their careers, but these estimates are not statistically significant at traditional levels. At minimum, though, this estimated positive correlation provides suggestive evidence that regression to the mean in observation ratings is not driving our results.

On the other hand, the positive relationship between initial effectiveness and early growth does not to persist for the second spline; the correlation between initial effectiveness and growth later in the career becomes negatively correlated ($r = -0.28$). This suggests that teachers who perform well in their first years of experience have lower growth rates later in their careers,

Table 4
Correlations Between Intercepts and Slope Terms

	Initial Effectiveness	Growth 0 to 2 Years of Experience
Observation ratings		
Growth 0 to 2 years of experience	.2826	—
Growth 3 to 5 years of experience	–.2778	.2047
Value-added measures		
Growth 0 to 2 years of experience	.1019	—
Growth 3 to 5 years of experience	–.7040**	–.2568

*$p < .05$.

an observation that could partially be explained by the regression to the mean argument for teacher effectiveness (Atteberry et al., 2015).[3]

The correlation patterns for VAMs are somewhat different. We find that initial effectiveness on VAMs is weakly correlated with growth early in the teaching career ($r = 0.10$) and is negatively correlated with growth later in the career ($r = –0.70$).[4]

Finally, we note that growth early in the career is negatively correlated with growth later in the career for observation ratings ($r = –0.28$) and for VAMs ($r = –0.70$). This result is in line with our prior results suggesting that the growth rates taper off with experience.

RQ 3: Which Program-Level Features Predict Graduates' Initial Effectiveness and Growth During Their Early Teaching Careers? Do the Same Features That Predict Initial Effectiveness Also Predict Later Growth?

Documenting the existence of significant differences across TPPs in the improvement of graduates' effectiveness over time highlights an important dimension of TPP effectiveness. However, policymakers and practitioners are likely interested in understanding which program features produce better initial effectiveness and more substantial growth. While our data cannot provide causal evidence, we provide some suggestive analyses using four key focal features: license type, program level, clinical placement type, and student teaching length. We view these as exploratory descriptive analyses aiming to provide insight into whether and how preparation features can be linked to both initial teaching effectiveness and growth over time; additionally, program features that are predictive of either or both should likely be considered further in future research using causal approaches and more comprehensive data on preparation features.

We estimate the relationships between different TPP features and graduates' initial effectiveness and growth by including our four focal features of TPPs in our preferred models, including the main effect of each feature and

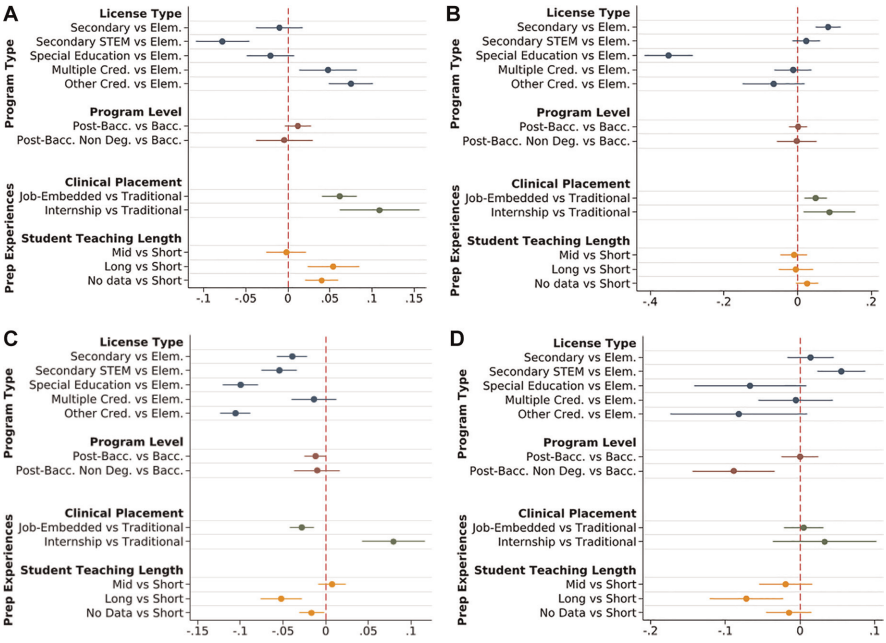

Figure 2. **Relationship between features of teacher preparation programs and growth estimates: (a) observation ratings: initial effectiveness, (b) teacher value-added measures: Initial effectiveness, (c) observation ratings: growth estimate 0-2 Years, and (d) teacher value-added measures: growth estimate 0-2 years.**

interacting it with our teacher experience predictors. In Figure 2, we report results for initial effectiveness and early-career (Years 0–2) growth trajectories; see Appendix Table 1, available online, for estimates from results for later (Years 3–5) growth trajectories. We estimated models with each feature separately and with all features included together (implicitly controlling for each feature). Because results were quite similar (see Appendix Table 1), we report here only on estimates from models where features are entered separately.

Program Type (Endorsement Area; Graduate/Undergraduate)

For observation ratings, we observe variation in initial effectiveness depending on the endorsements that graduates receive. Secondary STEM and special education teachers receive lower initial observation ratings than elementary teachers (−0.078 and −0.021 *SD*, respectively), while multiple and "other" credential teachers receive higher initial ratings (0.048 and 0.075 *SD*, respectively). However, elementary teachers exhibit greater

subsequent growth when compared to teachers in all other endorsement areas (all coefficients are negative, between −.014 and −.106, suggesting that elementary teachers, as the leave out group, has higher growth rates).

On VAMs, graduates with secondary (non-STEM) endorsements have higher levels of initial effectiveness than graduates with elementary endorsements (0.083 *SD*), while graduates with special education endorsements have initially lower levels (−0.350 *SD*). Graduates from all nonelementary endorsements have statistically similar rates of initial growth (i.e., none of the differences are statistically significant), except for secondary STEM teachers who grow at faster rates (0.055 *SD*, $p < .001$).

We find no differences in initial observation ratings or value-added between teachers who complete postbaccalaureate (degree or no degree) programs and teachers who complete 4-year baccalaureate programs. We find no differences in subsequent growth trajectories for observation ratings between these groups, although graduates from postbaccalaureate (no degree) programs have subsequently lower growth rates on VAMs.

Clinical Experiences

PCs who complete internship or job-embedded placements initially outperform PCs who complete traditional student teaching placements on both observation ratings (0.109 for internship; 0.062 for job-embedded) and VAMs (0.087 for internship; 0.049 for job-embedded). Subsequently, PCs who complete internship placements also demonstrate faster rates of growth (particularly on observation ratings by 0.079 *SD*) than PCs who complete traditional student teaching placements, while graduates from job-embedded programs grow at somewhat lower rates than graduates from traditional programs (−0.028 *SD* for observation ratings and 0.005 *SD* for VAM).

Graduates from TPPs that have longer clinical experiences (i.e., in the upper third of duration) tend to receive higher initial observation ratings (by 0.054 *SD*) but not VAMs (by −0.005 *SD*) than graduates who complete shorter placements. Graduates who completed the longest clinical placements also tend to have slower, subsequent growth rates in observation ratings and VAMs (−0.052 and −0.072, respectively).

Discussion

We highlight four key findings in this article. First, there is significant between-TPP variation in graduates' initial levels of instructional effectiveness in both classroom observation ratings and teacher VAMs. Only a small portion of the total variance in TTP graduates' effectiveness is explained by differences between TPPs (3.4% for initial observation ratings and 1.4% for initial VAMs). These estimates are similar in magnitude to the amount of variance explained by school districts in the state and to what others have reported for the differences among TPPs in other labor markets (Boyd et al., 2009;

Constantine et al., 2009; Darling-Hammond et al., 2005; Glazerman et al., 2006; Goldhaber et al., 2013; Henry, Purtell, et al., 2014; Koedel et al., 2015). Though TPPs explain only a small amount of the variation in the outcomes, average differences between some TPPs on these outcomes are nonetheless meaningful. The difference in outcomes for teachers who graduate from programs in the top decile and the statewide average is about 0.15 standard deviations for both observation ratings and VAMs, which is roughly half the growth that new teachers experience in their first year. In other words, compared to graduating from an average TPP, we find that graduating from the most effective TPPs in the state is comparable to about 4 and a half months of teaching experience.

Second, we also find significant and meaningful differences between TPPs in terms of graduates' growth in instructional effectiveness early in their careers. While most TPP evaluation policies focus only on initial effectiveness, there are important dynamics across programs in how teaching effectiveness develops over time. This has substantial implications for considering the overall effectiveness of different programs. And it suggests that differences in program design might contribute to these patterns, for example, if some programs specialize in preparing teachers to perform effectively out of the gate, while others prepare teachers to learn more effectively on the job.[5] Moreover, estimates of program graduates' initial effectiveness is only weekly correlated with estimates of subsequent growth. In other words, programs that seem to promote initial effectiveness are not necessarily those where teachers improve more early in their careers.

Third, finding that some programs excel in graduating teachers who are effective initially, while others excel in graduating teachers who grow more rapidly early in their careers, led us to consider whether certain program features may be related to initial effectiveness while others may be related to growth, or whether there may be some features that predict both initial effectiveness and growth. The analyses linking features of preparation and graduates' initial effectiveness and later growth are meant to serve as a methodological exploration and a proof of concept that our modelling approach could identify meaningful differences among TPPs, not as definitive evidence of the impacts of specific features. Consistent with past work (e.g., Ronfeldt et al, 2020), none of the TPP features we studied were consistently related to the outcomes in our analyses. If anything, we find suggestive evidence that internship programs are more effective in promoting both initial effectiveness and growth on observation ratings and VAMs. By and large, we find that some features predict initial effectiveness, while others tend to predict growth—again highlighting the importance of exploring both dimensions of effectiveness. These results suggest that the differences we observe among TPPs could be driven, at least in part, by observed differences in program type or preparation experiences.

Fourth, our analyses of program features, though correlational, identify aspects of preparation to investigate further—ideally with experimental or causal inference methodologies—as aspects of preparation that potentially cause graduates to be more instructionally effective and/or grow more quickly. Regarding student teaching duration, a recent review of the quantitative evidence suggests that prior studies have found longer student teaching to be associated with graduates feeling better prepared to teach but not more instructionally effective (on observation ratings and VAMs) (Ronfeldt, 2021). Our results are somewhat at odds with the latter, where we find mixed evidence: Longer duration predicts higher initial observation ratings but lower initial VAMs and subsequent growth on VAMs.

Our results regarding clinical placement types are consistent with recent evidence suggesting that quality and kinds of clinical placements are associated with various workforce outcomes. Relative to traditional student teaching placements, we find internship placements to be associated with higher initial effectiveness and subsequent growth (on both effectiveness measures). Given internship placements are typically associated with residency model programs, these results are perhaps consistent with recent evidence that this model has promise, though more research is needed to determine if the effects are truly due to the placement type (rather than other dimensions of these programs) and whether the effects are truly causal.

Regarding endorsement type, graduates endorsed in special education, as compared with elementary endorsements, had lower initial observation ratings, lower growth on observation ratings, and lower initial VAMs. Though these results are consistent with prior research in the same state (Bardelli & Ronfeldt, 2020), these differences may not reflect differences in preparation. For example, prior evidence indicates that lower observation ratings and VAMs for teachers of many students with special needs might be lower not because they are less-effective teachers but because these measures are sensitive to the populations of students with whom they work (Buzick & Jones, 2015; Campbell & Ronfeldt, 2018; Jones & Brownell, 2014).

We intend this study to illustrate the value of examining growth as well as initial effectiveness, and to highlight that programs and program features related to initial effectiveness may not be the same programs and features that predict future growth. However, our work has several limitations. First, it is purely correlational, and we cannot estimate the causal effect of individual programs—or program features—on outcomes. As discussed above, various forms of selection may explain these differences across programs and features, including the selection of prospective teachers to TPPs, clinical placements, and employment schools. As a result, we cannot assume either differences between TPPs or the relationships between TPP features and outcomes to be causal in nature. Second, we focus on a very narrow set of instructional effectiveness outcomes. While we extend on past studies by exploring observation ratings, it is likely that TPPs influence their graduates on facets of teaching practice not

captured by these two measures, such as effectiveness is supporting student socio-emotional development or anti-racist teaching practices. Moreover, prior studies have found that features predicting teacher retention are often not the same as those predicting instructional effectiveness.

Finally, while this study suggests that TPPs should attend to growth on observation ratings and VAMs in addition to initial effectiveness, we recognize that instructional effectiveness outcomes collected through teacher evaluation provide a narrow lens through which to evaluate TPPs. Arguably, the goal of teacher preparation is to provide well-prepared teachers who support the learning of diverse students from a wide range of backgrounds. Especially given the need to prepare candidates for diverse contexts and student populations, focusing on a narrow set of instructional effectiveness outcomes might overlook other outcomes that TPPs should focus on, such as an attention to equity-oriented teaching or knowledge of and skill with local communities. Future work should investigate whether and to what extent TPP graduates' initial level and future growth on these noninstructional effectiveness measures vary across TPPs and are associated with different features of preparation.

Conclusion

Despite these limitations, we think these results have important lessons for policymakers and suggest valuable directions for future study. Most importantly, these findings suggest that policymakers should consider both initial preparation and later growth, rather than focusing solely on average initial effectiveness as is the case in many evaluation systems. The ultimate goal of TPP evaluation policies should be to identify programs and practices that lead to improved outcomes for students and their teachers. The same is true for local decisionmakers who are deciding which teachers to hire. Thus, any efforts to evaluate TPPs based on the effectiveness of their graduates should account for improvement as teachers gain experience in the classroom. Hiring teachers who are more effective in the first year but who fail to develop their practice over time (or who leave the classroom), is not a sustainable approach for staffing schools successfully. Focusing only on initial effectiveness may obscure programs that set teachers up for longer-term success. A more nuanced evaluation policy would likely provide more comprehensive feedback to TPPs to inform program leaders about how to improve.

Our modeling approach and results also suggest that researchers might move beyond focusing on graduates' average levels of effectiveness and consider initial effectiveness and subsequent growth together in efforts to evaluate teacher education program features. We think that future studies should aim to use experimental or causal inference methodologies to determine whether these correlational patterns are causal in nature or not and to consider a wider range of program features as well as outcomes.

While a causal analysis of the sources of these differences is beyond the scope of the present study and an area in need of future research, we propose three possible explanations for future studies to interrogate. First, having lower initial effectiveness may leave more room for subsequent growth; and, likewise, having higher initial effectiveness may leave less room for subsequent growth. This explanation is consistent with prior literature showing regression to the mean among instructional effectiveness measures like those we study here (Atteberry et al., 2015). Relatedly, this could suggest that candidates who enter with weaker preparation must depend upon on-the-job growth to make up for shortcomings of initial preparation. However, both regression to the mean or compensatory growth would likely yield faster early growth (first spline) for traditional graduates so these explanations seem unlikely to account for why we only observe faster later growth (second spline). Second, context matters. Perhaps preparation programs that focus heavily on preparing students to teach in a given context may produce graduates with greater initial effectiveness. But as context changes over time (e.g., as new curricular materials are brought in), teachers who are prepared with a narrow focus on a specific district placement may not have the same skills to learn and adapt. Finally, traditional route programs often emphasize theory and reflection (including in coursework), which is based upon the argument that these will better prepare graduates to be "students of teaching" (Dewey, 1904) who will be better equipped to learn and grow on the job (Kennedy, 1987). It is possible that these emphases explain the relationship we observe between traditional student teaching placements and later growth. This argument is consistent with prior qualitative research suggesting that the benefits of theory learned during initial preparation may not manifest initially but instead during later years (Grossman & Richert, 1988).

Supplemental Material

Supplemental material for this article is available online.

Notes

We are grateful to Andrew Grogan-Kaylor, Kevin Schaaf, Paul von Hippel, and reviewers at the Tennessee Education Research Alliance (TERA) for feedback on early drafts of this manuscript. We are also grateful for the feedback from participants of the 2019 and 2021 American Educational Research Association (AERA) annual meetings and of the Causal Inference in Educational Research Seminary (CIERS) at the University of Michigan. All omissions and errors are our own. Emanuele Bardelli received support from the Institute of Education Sciences, U.S. Department of Education (PR/Award # R305B1170015) to complete this work as part of a predoctoral training fellowship. The Advanced Research Computing at the University of Michigan–Ann Arbor also in part supported this work by providing computational resources and services.

[1]Note that both observation ratings and VAMs are negative. These outcomes are standardized on the universe of teachers within each year. A negative value means that teachers

in our sample receive below average observation ratings and VAMs. This is expected as these teachers are in the first years of their teaching careers.

[2]See, for example, https://web.archive.org/web/20090413150009/http://tn.gov/education/lic/license_types.shtml or https://web.archive.org/web/20060216140157/http://www.tnt2t.com/.

[3]One point to make about regression to the mean is that we report estimates for TPPs. While regression to the mean can influence the results from an individual teacher, it is less likely (and maybe implausible) that all graduates from the same program experience the same regression to the mean pattern across all observation years. That is, regression to the mean likely biases (possibly upwards) the within-program variation in teacher scores; however, it is less likely that it explains all the between-program differences that we observe. For example, for regression to the mean to explain all our between-program differences, we need to observe graduates from the same programs to all draw a high observation rating during their first year of teaching and then all draw low the following year. This process needs to repeat for all programs and all observation years, on average. While it is possible for a purely stochastic process to produce these patterns, it is safe to assume that its probability is very close to zero in practice.

[4]It is worth noting here that the results that we observe in Appendix Figure 3 and the correlation patterns for VAMs do not suggest regression to the mean, in that we have significant heterogeneity in initial effectiveness (Year 0) and no significant heterogeneity in early growth (Years 0–2). We interpret these results to mean that the observed differences between TPPs in initial effectiveness are carried over during their early career period. Were regression to the mean to exist, we would expect (a) there to be significant differences between TPPs during early growth (Years 0–2) and (b) the correlations between initial effectiveness and early growth to be negative; however, we observe neither of these to be true. More formally, the estimates for differences between TPPs in VAMs growth during the early career (i.e., Years 0–2) are not significant. That is, all teachers' VAMs grow on average about 0.123 standard deviation units during this time and that there are no statistical differences between programs in these growth rates. This nonsignificant heterogeneity in early growth rates suggests that the initial differences between TPPs in VAMs largely maintain during the first three years of teaching, leading to the weak correlation patterns that we observe between initial VAMs effectiveness and early growth.

[5]Alternatively, faster on-the-job growth among graduates from some TPPs could reflect shortcomings of the preparation provided by these TPPs rather than some kind of specialized preparation for how to effectively learn from experience; in other words, inadequate initial preparation may have required that graduates learn more rapidly on the job to compensate. We suspect this explanation is, on average, unlikely given that we observe initial effectiveness to be positively associated with early growth. A compensatory explanation would likely yield a negative correlation.

References

Atteberry, A., Loeb, S., & Wyckoff, J. (2015). Do first impressions matter? Predicting early career teacher effectiveness. *AERA Open*, *1*(4), 2332858415607834. https://doi.org/10.1177/2332858415607834

Bacher-Hicks, A., Chin, M. J., Kane, T. J., & Staiger, D. O. (2017). *An evaluation of bias in three measures of teacher quality: Value-added, classroom observations, and student surveys* (No. w23478). National Bureau of Economic Research.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., & Shepard, L. A. (2010). *Problems with the Use of Student Test Scores to Evaluate Teachers* (EPI Briefing Paper #278). Economic Policy Institute.

Bardelli, E., & Ronfeldt, M. (2020). *Workforce outcomes of program completers in high needs areas* (Working Paper No. 2020–01). Tennessee Education Research

Alliance, Vanderbilt University. https://peabody.vanderbilt.edu/TERA/files/TERA_Working_Paper_2020-01.pdf

Bartanen, B., & Kwok, A. (2020). *Pre-service teacher quality and workforce entry* (EdWorkingPaper 20-223). Annenberg Institute at Brown University. https://www.edworkingpapers.com/ai20-223

Bastian, K. C., Patterson, K. M., & Pan, Y. (2018). Evaluating teacher preparation programs with teacher evaluation ratings: Implications for program accountability and improvement. *Journal of Teacher Education*, *69*(5), 429–447. https://doi.org/10.1177/0022487117718182

Bitler, M., Corcoran, S. P., Domina, T., & Penner, E. K. (2021). Teacher effects on student achievement and height: A cautionary tale. *Journal of Research on Educational Effectiveness*, *14*(4), 900–924.

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. https://doi.org/10.3102/0162373709353129

Boyd, D. J., Lankford, H., Loeb, S., & Wyckoff, J. (2005). The draw of home: How teachers' preferences for proximity disadvantage urban schools. *Journal of Policy Analysis & Management*, *24*(1), 113–132. https://doi.org/10.1002/pam.20072

Buzick, H. M., & Jones, N. D. (2015). Using test scores from students with disabilities in teacher evaluation. *Educational Measurement: Issues and Practice*, *34*(3), 28–38. https://doi.org/10.1111/emip.12076

California State University (CSU). (2002). *First system wide evaluation of teacher education programs in the California State University: Summary report*.

Campbell, S. L., & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal, 55*(6), 1233–1267. https://doi.org/10/gd32fh

Cannata, M. (2011). The role of social networks in the teacher job search process. *Elementary School Journal*, *111* (3), 1–24.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593–2632.

Cochran-Smith, M., & Lytle, S. L. (1993). *Inside/outside: Teacher research and knowledge*. Teachers College Press. https://market.android.com/details?id=book-H4uwnL1IPvUC

Constantine, J., Player, D., Silva, T., Hallgren, K., Grider, M., & Deke, J. (2009). *An evaluation of teachers trained through different routes to certification* (NCEE 2009-4043). National Center for Education Evaluation and Regional Assistance. https://eric.ed.gov/?id=ED504313

Cowan, J., Goldhaber, D., & Theobald, R. (2017). *Massachusetts educator preparation and licensure*. American Institutes for Research. https://www.doe.mass.edu/research/reports/2017/05EdPrep-Year1Report.pdf

Darling-Hammond, L., Holtzman, D. J., Gatlin, S. J., & Vasquez Heilig, J. (2005). Does teacher preparation matter? Evidence about teacher certification, Teach for America, and teacher effectiveness. *Education Policy Analysis Archives*, *13*(42). http://www.redalyc.org/html/2750/275020513042/

Dewey, J. (1904). The relation of theory to practice in education. In C. A. McMurry (Ed.), *The third yearbook of the National Society for the Scientific Study of Education*. University of Chicago Press.

Giersch, J., & Dong, C. (2018). Principals' preferences when hiring teachers: A conjoint experiment. *Journal of Educational Administration*, *56*(4), 429–444. https://doi.org/10.1108/JEA-06-2017-0074

Glazerman, S., Mayer, D., & Decker, P. (2006). Alternative routes to teaching: The impacts of Teach for America on student achievement and other outcomes. *Journal of Policy Analysis and Management*, *25*(1), 75–96. https://doi.org/10.1002/pam.20157

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

Goldhaber, D., & Ronfeldt, M. (2020). Towards causal evidence on effective teacher preparation. In J. E. Carinci, K. Jackson, & S. J. Meyer (Eds.), *Linking teacher preparation program design and implementation to outcomes for teachers and students* (pp. 211–236). IAP.

Grossman, P. L., & Loeb, S. (2008). *Alternative routes to teaching: Mapping the new landscape of teacher education*. Harvard Education Press.

Grossman, P. L., & Richert, A. E. (1988). Unacknowledged knowledge growth: A reexamination of the effects of teacher education. *Teaching and Teacher Education*, *4*(1), 53–62.

Harris, D. N., Rutledge, S. A., Ingle, W. K., & Thompson, C. C. (2010). Mix and match: What principals really look for when hiring teachers. *Education Finance and Policy*, *5*(2), 228–246. https://doi.org/10.1162/edfp.2010.5.2.5205

Harris, D. N., & Sass, T. R. (2014). Skills, productivity and the evaluation of teacher performance. *Economics of Education Review*, *40*, 183–204. https://doi.org/10.1016/j.econedurev.2014.03.002

Henry, G. T., Bastian, K. C., Fortner, C. K., Kershaw, D. C., Purtell, K. M., Thompson, C. L., & Zulli, R. A. (2014). Teacher preparation policies and their effects on student achievement. *Education Finance and Policy*, *9*(3), 264–303. https://doi.org/10.1162/EDFP_a_00134

Henry, G. T., Fortner, C. K., & Bastian, K. C. (2012). The effects of experience and attrition for novice high-school science and mathematics teachers. *Science*, *335*(6072), 1118–1121. https://doi.org/10.1126/science.1215343

Henry, G. T., Purtell, K. M., Bastian, K. C., Fortner, C. K., Thompson, C. L., Campbell, S. L., & Patterson, K. M. (2014). The effects of teacher entry portals on student achievement. *Journal of Teacher Education*, *65*(1), 7–23. https://doi.org/10.1177/0022487113503871

Jabbar, H., Cannata, M., Germain, E., & Castro, A. (2020). It's who you know: The role of social networks in a changing labor market. *American Educational Research Journal*, *57*(4), 1485–1524. https://doi.org/10.3102/0002831219879092

Jacob, B. A., Rockoff, J. E., Taylor, E. S., Lindy, B., & Rosen, R. (2018). Teacher applicant hiring and teacher performance: Evidence from DC public schools. *Journal of Public Economics, 166*, 81–97. https://doi.org/10.1016/j.jpubeco.2018.08.011

Jones, N. D., & Brownell, M. T. (2014). Examining the use of classroom observations in the evaluation of special education teachers. *Assessment for Effectiveness Intervention*, *39*(2), 112–124. https://doi.org/10.1177/1534508413514103

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

Kennedy, M. M. (1987). *Inexact sciences: Professional education and the development of expertise* (Issue Paper 87-2). National Center for Research on Teacher Education.

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

Kraft, M. A., Brunner, E. J., Dougherty, S. M., & Schwegman, D. J. (2020). Teacher accountability reforms and the supply and quality of new teachers. *Journal of Public Economics, 188*, 1–24. https://doi.org/10.1016/j.jpubeco.2020.104212

Kraft, M. A., & Papay, J. P. (2014). Can professional environments in schools promote teacher development? Explaining heterogeneity in returns to teaching experience. *Educational Evaluation and Policy Analysis, 36*(4), 476–500. https://doi.org/10.3102/0162373713519496

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. https://doi.org/10.1162/EDFP_a_00110

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

Papay, J. P., & Laski, M. E. (2020). Understanding the Dynamics of Teacher Productivity Development: Evidence on Teacher Improvement in Tennessee [Working paper].

Papay, J. P., West, M. R., Fullerton, J. B., & Kane, T. J. (2012). Does an urban teacher residency increase student achievement? Early evidence from Boston. *Educational Evaluation and Policy Analysis*. https://doi.org/10.3102/0162373712454328

Preston, C. (2017). University-based teacher preparation and middle grades teacher effectiveness. *Journal of Teacher Education*, *68*(1), 102–116. https://doi.org/10.1177/0022487116660151

Reininger, M. (2012). Hometown disadvantage? It depends on where you're from: Teachers' location preferences and the implications for staffing schools. *Educational Evaluation and Policy Analysis, 34*(2), 127–145. https://doi.org/10.3102/0162373711420864

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economics Review, 94*(2), 247–252.

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy, 6*(1), 43–74. https://doi.org/10.1162/EDFP_a_00022

Ronfeldt, M. (2012). Where should student teachers learn to teach? *Educational Evaluation and Policy Analysis*, *34*(1), 3–26. https://doi.org/10.3102/0162373711420865

Ronfeldt, M. (2015). Field placement schools and instructional effectiveness. *Journal of Teacher Education*, *66*(4), 304–320. https://doi.org/10.1177/0022487115592463

Ronfeldt, M. (2021). Links among teacher preparation, retention, and teaching effectiveness. Evaluating and Improving Teacher Preparation Programs. National Academy of Education Committee on Evaluating and Improving Teacher Preparation Programs. National Academy of Education. https://doi.org/10.31094/2021/3/1

Ronfeldt, M., & Campbell, S. L. (2016). Evaluating teacher preparation using graduates' observational ratings. *Educational Evaluation and Policy Analysis*, *38*(4), 603–625.

Ronfeldt, M., Matsko, K. K., Greene Nolan, H., & Reininger, M. (2020). Three different measures of graduates' instructional readiness and the features of preservice preparation that predict them. *Journal of Teacher Education*, Online First. https://doi.org/10.1177/0022487120919753.

Ronfeldt, M., & Reininger, M. (2012). More or better student teaching? *Teaching and Teacher Education*, *28*(8), 1091–1106. https://doi.org/10.1016/j.tate.2012.06.003

Ronfeldt, M., Schwartz, N., & Jacob, B. (2014). Does pre-service preparation matter? Examining an old question in new ways. *Teachers College Record*, *116*(10), 1–46.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *Quarterly Journal of Economics*, *125*(1), 175–214.

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action* (Vol. 5126). Basic Books.

von Hippel, P. T., & Bellows, L. (2018). How much does teacher quality vary across teacher preparation programs? Reanalyses from six states. *Economics of Education Review*, *64*, 298–312. https://doi.org/10.1016/j.econedurev.2018.01.005

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

Vosters, K. N., Guranio, C. M., & Wooldridge, J. M. (2018). *Understanding and evaluating the SAS EVAAS univariate response model (URM) for measuring teacher effectiveness* (UNC Charlotte Economics Working Paper Series). https://belkcollegeofbusiness.uncc.edu/economic-working-papers/wp-content/uploads/sites/850/2018/06/wp2018-001.pdf

Manuscript received August 27, 2021
Final revision received August 17, 2022
Accepted October 6, 2022