ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 21-450

Teacher Preparation Programs and Graduates' Growth in Instructional Effectiveness

Emanuele Bardelli
University of Michigan

Matthew Ronfeldt
University of Michigan

John Papay
Brown University

Many prior studies have examined whether there are average differences in levels of teaching effectiveness among graduates from different teacher preparation programs (TPPs); other studies have investigated which features of preparation predict graduates' average levels of teaching effectiveness. This is the first study to examine whether there are average differences between TPPs in terms of graduates' average growth, rather than levels, in teaching effectiveness, and to consider which features predict this growth. Examining all graduates from Tennessee TPPs from 2010 to 2018, we find meaningful differences between TPPs in terms of both levels and growth in teaching effectiveness. We also find that different TPP features, including areas of endorsement, program type, clinical placement type and length, program size, and faculty composition explain part of these differences. Yet, the features that predict initial teaching effectiveness are not the same features that predict growth.

VERSION: August 2021

Suggested citation: Bardelli, Emanuele, Matthew Ronfeldt, and John Papay. (2021). Teacher Preparation Programs and Graduates' Growth in Instructional Effectiveness. (EdWorkingPaper: 21-450). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/120j-qb52

Teacher Preparation Programs and Graduates' Growth in Instructional Effectiveness

Emanuele Bardelli

Matthew Ronfeldt

University of Michigan

John Papay

Brown University

Abstract: Many prior studies have examined whether there are average differences in levels of teaching effectiveness among graduates from different teacher preparation programs (TPPs); other studies have investigated which features of preparation predict graduates' average levels of teaching effectiveness. This is the first study to examine whether there are average differences between TPPs in terms of graduates' average growth, rather than levels, in teaching effectiveness, and to consider which features predict this growth. Examining all graduates from Tennessee TPPs from 2010 to 2018, we find meaningful differences between TPPs in terms of both levels and growth in teaching effectiveness. We also find that different TPP features, including areas of endorsement, program type, clinical placement type and length, program size, and faculty composition explain part of these differences. Yet, the features that predict initial teaching effectiveness are not the same features that predict growth.

Acknowledgments: 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.

Teacher Preparation Programs and Graduates' Growth in Instructional Effectiveness

Introduction

Pre-service teacher education is the main entryway into teaching. It is not surprising, then, that researchers and policymakers have been trying to develop ways to assess teacher preparation programs (TPPs) using graduates' workforce outcomes. The general theory behind these policies is that the curriculum, pedagogical orientation, and clinical experiences among TPPs likely affect TPP graduates' workforce outcomes but vary in quality across providers. For example, the Council of Chief State School Officers (2019) has advocated for the development and strengthening of TPP evaluation, while the Council for the Accreditation of Educator Preparation (CAEP)'s accreditation requirements include a program impact component.

Researchers, on the other hand, have found mixed evidence to support the use of graduates' value-added to student test scores (VAM) 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 fact, von Hippel and Bellows (2018) showed that most of the observed differences among TPP graduates' VAMs 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. Ronfeldt and Campbell (2016) and Bastian, Patterson, and Pan (2018) found significant variation between TPPs in terms of their graduates' average observation ratings during their early careers.

These analyses have focused on average levels of VAMs or instructional practice for early-career teachers from different preparation programs. However, Goldhaber, Liddle, & Theobald (2013) argue that TPP's impacts on teacher practice are long-lasting, continuing throughout the first decade in the classroom. Furthermore, we know that teacher practice and effectiveness in raising student test scores develop rapidly over the first few years in the classroom (Rockoff, 2004; Harris & Sass, 2014; Atteberry et al., 2015; Papay & Kraft, 2015).

As a result, TPPs may not only affect teachers' *level* of effectiveness but also the rate at which they *improve* their practice. For example, successful TPPs could prepare not only initially effective teachers but also reflective professionals that can learn while on the job. 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 (Harris et al., 2010; Giersch & Dong, 2018).

In this paper, we use data from TPP graduates in Tennessee to estimate initial performance and growth in performance over time for graduates from different programs. We focus on VAM scores and direct assessments of instructional performance using the state's classroom observation system. With these measures, we find significant variation among TPPs in the state in both graduates' initial effectiveness and 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 capture unique aspects of teacher preparation.

Finding substantial between-TPP variation in initial effectiveness and subsequent growth led us to wonder which preparation features might contribute to the observed variation. As a number of scholars have explored how features of preparation, such as certification area, program type, and student teaching duration, relate to graduates' instructional effectiveness, it is not clear that the same features of preparation would relate to both initial effectiveness and the propensity for teachers to improve over time. For example, we may see a trade-off in curricular emphasis between a focus on being prepared on day one and an orientation towards reflective praxis. We find that internships and job-embedded programs have graduates who are initially more effective but have different improvement trajectories over time, with graduates from job-embedded programs having slower improvement trajectories than graduates from internship programs. Moreover, we find that graduates from programs with longer student teaching placements have higher initial effectiveness scores but growth at lower rates than other graduates.

These results highlight the work that TPPs do with their students beyond preparing students to enter the classroom as novice teachers; as such, it helps to articulate an additional dimension of teacher preparation—the propensity to support career growth over time. These results also suggest that the current TPP evaluation systems might miss an important component of teacher preparation.

Literature Review

Differences in the Evaluation Outcomes for Graduates of Different TPPs

Individual TPPs have unique features and characteristics—curriculum, course sequence and structures, clinical placements—that they purport will best support candidate learning and ensure that their graduates are effective teachers. Only within the last fifteen years, however, it has become possible to assess whether these differences among TPPs relate to differences in outcomes for their graduates. This prior research has found mixed results.

Several studies have suggested that there are average differences between programs in terms of graduates' instructional effectiveness, as measured by contributions to student test scores. Boyd et al. (2009) was among the first papers to estimate differences across TPPs, finding that, in NYC schools, "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) and about "a two-standard error difference between the higher and lower value-added programs" (p. 429). Expanding on this work, Goldhaber et al. (2013) found more limited variation for graduates from different TPPs in Washington state. They reported that TPP's 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 to 15 years. This suggests that TPPs have long-lasting impacts on their graduates' effectiveness.

Studies have found more robust differences between TPPs when exploring graduates' instructional practice as measured by classroom observation ratings. Using data from Tennessee, Ronfeldt and Campbell (2016) found significant variation in graduates' observation ratings across TPPs, reporting that "institutions explained about 2% of variation in graduates' [teacher

VAM] scores and 4% of variation in [observation ratings]; programs explained about 4% of variation on both outcomes" (p. 610). Bastian, Patterson, and Pan (2018) found similar results in North Carolina, concluding that "TPPs are significantly associated with the evaluation ratings of their graduates" (p. 442).

Other studies have raised methodological challenges in estimating TPP effects, suggesting that there might be no actual differences between TPPs in terms of graduates' average instructional effectiveness. These studies argued that studies estimating TPP effects are unable to confidently distinguish signal from noise in their estimation of program effects on graduates' instructional effectiveness. Koedel et al. (2015) critiqued this prior work for not taking into account teacher sampling and reported that most of the variation in teacher evaluation scores is explained by differences within programs rather than differences between programs. They used data from 24 TPPs in Missouri and showed that "cross program differences explain no more than 3.4 percent of the total variance in teacher value added in any model, and as little as 1.6 percent" (p. 520).

Von Hippel et al. (2016) made these challenges more explicit, using meta-analytic methods to show that most of the variation in graduates' teacher evaluation scores for Texas TPPs could be explained by random measurement error across TPPs. 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) re-analyzed data from NYC, Washington state, Missouri, and Florida, and again found little evidence of differences across TPPs beyond expected random noise.

Mihaly et al. (2013) raised an additional concern with implications for how to address the clustering of program graduates in certain schools. They argue 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 percent of the within-TPP variability in teacher evaluation scores, leading to the TPP variance in models with fixed effects to be almost twice the variance in models without TPP fixed effects.

These more critical 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.

TPP Effects on Teachers' Effectiveness at Entry vs. 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 performance 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 evolve over time.

Program Features that Correlate with Graduates' Outcomes

To the extent that TPPs differ in how well they support the development of teacher effectiveness throughout the career, we might be interested in exploring what features of TPP programs correlate with these differences. Here, there is a long tradition in examining how differences across programs in their practices, structures, and norms explain differences in average workforce outcomes for TPP graduates.

In fact, Boyd et al. (2009) reported on the relationship between program features and TPP graduates' outcomes. They included in their analyses features such as whether a program required a capstone project, the level of oversight of graduates' field placement experiences, content knowledge requirements measured as the number of courses in mathematics and ELA that graduates completed, and percentage of tenure-track faculty as a proxy for program stability and institution commitment to teacher preparation. They found some variation across cohorts in the effects of these features over time but, in general, found positive effects for all variables. This work suggests that these features, which we can see as the TPP building blocks, can explain the relationship between TPP and graduates' workforce outcomes.

In the sections below, we review specific literature on particular features that have been a focus for researchers and that we include in our analyses.

Certification Area

Though there is substantial literature examining whether early career teachers who completed certification are more or less effective than teachers who are uncertified or alternatively certified, less research has investigated whether there are differences in instructional effectiveness between teachers from different certification/endorsement areas, e.g., secondary mathematics, elementary, special education. Most prior analyses have focused on either STEM or SpEd certification.

STEM Certification. 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, SpEd, and bilingual/ESL endorsements. The only case where HNEA outperformed non-HNEA teachers was that teachers with STEM endorsements had greater mathematics VAMs than non-HNEA teachers. 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 teachers received significantly lower observation ratings than non-HNEA peers.

Henry, Fortner, and Bastian (2012) is the only prior study, to our knowledge, that has considered differences between certification area in terms of graduates' growth in, rather than level of, instructional effectiveness. Specifically, they find that VAM scores of science teachers grow at about double the rate of VAM scores of mathematics teachers and about four times the rate of VAM scores of non-STEM teachers.

SpEd Certification. Bardelli and Ronfeldt (2020) found that early career teachers with SpEd endorsements have lower observation ratings and VAMs than non-HNEA teachers. Similarly, Ronfeldt (2015) finds that teachers with SpEd certification have lower average mathematics VAMs than teachers from other certification areas. In Florida, Feng and Sass (2013) find that students with disabilities have greater achievement gains when taught by SpEd certified teachers as compared to teachers certified in other areas; however, students without disabilities do slightly worse. Buzick and Jones (2015) find that generally teachers' VAM scores are similar with and without special education students included, except in classrooms with high proportions of special education students–teachers in these classrooms have lower average VAM scores; though the authors suggest that including controls for SpEd students increase these teachers' VAM scores in ways that may mitigate concerns.

Other Certification. The only other HNEA that Bardelli and Ronfeldt (2020) investigated were bilingual/ESL endorsements. The authors found that teachers with bilingual/ESL certification had similar VAMs and observation ratings as teachers certified in non-HNEAs. Similarly, Master, Loeb, Whitney, and Wychoff (2016) reported that novice teachers with a bilingual/ESL endorsement have higher VAM scores than their peers without an endorsement. These effects appear to be concentrated with elementary school teachers.

Program Type

A number of studies have compared average teaching effectiveness of graduates from undergraduate versus graduate/postbac programs, showing mixed findings. Three studies considered both mathematics and ELA VAMs. Considering mathematics, Henry, Bastian et al. (2014) found that undergraduate programs outperformed graduate programs in terms of middle grades mathematics. 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) had access to VAMs in many 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 performance in other subject areas.

Program Size

We are aware of only one prior study that has considered whether program size is related to graduates' instructional effectiveness. Ronfeldt and Campbell (2016) compared programs in Tennessee in terms of average observation ratings among graduates. Because it was difficult to get good estimates for smaller programs, the authors combined graduates from all smaller programs across the state. On average, graduates from smaller programs received significantly lower observation evaluations than the mean of recent graduates across the state.

Percent Tenure Track Faculty

To our knowledge, only one study has tested whether graduates' teaching effectiveness is related to the proportion of tenure-line faculty in certification programs. Looking across all teacher education programs that supplied teachers to NYC schools, Boyd et al. (2009) examined a great number of features including percentage of certification program faculty that are tenure-lined which they thought to be "a potential proxy for program stability and the extent to which institutions value teacher preparation" (p. 430). They found that graduates from programs with a higher proportion of tenured faculty had significantly stronger achievement gains during the first year of inservice teaching.

Student Teaching Duration

A number of studies have found that graduates who complete longer student teaching experiences report feeling better prepared to teach (CSU, 2002 ; Ronfeldt et al., 2020; Ronfeldt et al., 2014) (Ronfeldt & Reininger, 2012) . Though it may be associated with feeling better prepared, there is little evidence that completing more student teaching is related to being more instructionally effective. Only one study considered graduates' observation ratings as the focal outcome and found it to be unrelated to the number of weeks of student teaching completed (Ronfeldt et al., 2020). Three other studies have linked measures of student teaching duration to graduates' VAMs. In terms of mathematics VAMs, one study found inconsistent results across model specifications (Boyd et al., 2009), another found no relationship (Ronfeldt, 2015), while the third found mixed but mostly negative associations (Preston, 2017). There were no significant relationships between ELA VAMs and student teaching duration across models and

studies, except for a significant, negative relationship in one of Preston's (2017) model specifications.

Clinical Placement Type

The US department of education has three categories for clinical placement experiences: (1) student teaching, (2) internship, and (3) job-embedded (see US Title II, Section 202 (d)(2) 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 are typically a full year in the same classroom co-teaching with the same cooperating/mentor teacher. Both *student teaching* and *internship* placements as pre-service 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 one is completing 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 and usually in the room 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 teaching effectiveness or 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 it is one distinguishing factor; namely, student teaching tends to be a semester, internships are typically a yearlong, and job-embedded placements are typically two 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. 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 non-residency 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. For example, the timing of coursework often differs between routes. In traditional programs, candidates typically complete most/all of their coursework prior to their student teaching semesters whereas individuals in other routes complete their coursework requirements while in their clinical placements. Additionally, the presence and role of the mentor/cooperating teacher also differs across routes.

Though on the face of it, our analysis of placement type can be considered a unique contribution of this study, it is important to consider results in light of the concerns raised above. In particular, any differences we observe between placement types are likely to be conflated with route of entry. Thus, we recommend caution in interpreting results.

Research Question

The research questions that guide this paper are:

RQ 1. How much of the variance in early career teachers' initial effectiveness and growth is explained by teacher preparation programs?
RQ 2. Are there differences between programs in terms of graduates' initial effectiveness and growth during their early teaching careers?
RQ 3. What program-level features predict graduates' initial effectiveness and growth during their early teaching careers?

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 evaluation outcomes. We draw information about TPP features—hours of clinical practice, total program size, and percent of faculty on the tenure track faculty—from the US Department of Education Title 2 dataset.

Sample

We focus on the universe of 43,917 TPP graduates in the state from 2010 to 2018, or about 30% of all teacher-year observations during this period. Thus, our analysis focuses on the performance of recent graduates from in-state TPPs.

We report summary statistics for our analytic sample in Table 1. Recent graduates from TN TPPs appear to perform similarly to out-of-state early career teachers on both teacher observation ratings and VAM scores[1]. Also, teachers prepared in-state do not differ from other teacher covariates from teachers prepared out-of-state.

Focal Outcomes

We focus on two outcomes to assess teacher effectiveness: (1) observation ratings of teachers and (2) value-added measures (VAM). These measures are collected as part of Tennessee's

[1] Note that both observation ratings and VAM scores 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 VAM scores. This is expected as these teachers are in the first years of their teaching careers.

teacher evaluation system and are used to calculate a summative evaluation score for each teacher in the state.

Observation ratings are based on the TEAM observation rubric. This rubric organizes twenty-four indicators into four domains: instruction, planning, environment, and professionalism. Each teacher in the state is evaluated 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. The state began its evaluation system during the 2011-2012 school year.

VAMs are available for a subset of teachers who teach tested grades and subjects—math, ELA, science, or social studies in grades 3-8 and most math and ELA teachers in high school. This includes roughly forty-four percent of the teachers who work in the state. These measures are calculated following the EVAAS model (Vosters et al., 2018) that seeks to isolate each teacher's contribution to students' test scores. We create average composite VAM scores for teachers who have scores in multiple tested subjects and standardize the estimates to have mean 0 and standard deviation of 1. VAMs are available starting from the 2010-2011 school year.

We finally standardize the observation ratings and VAMs within each school year using the state-wide sample of all teachers. This ensures that year-to-year changes in evaluation scores are taken out from the teacher-level scores by demeaning these scores within each year and that the evaluation scores are centered at the state-level mean. We interpret a teacher having a zero in evaluation scores as a teacher receiving an average evaluation score across school years and compared to all teachers in the state.

Teaching Experience

Years of teaching experience is our main explanatory variable of interest as we focus on modeling growth in evaluation scores over time. We measure teaching experience in several ways. First, we define experience as the number of years teaching since graduating from a TPP. We count as having zero years of experience the first year during which a program completer (PC) is reported as teaching in Tennessee. This school year can be either concurrent to the graduation year for PCs who complete an internship or job-embedded program or after the graduation year for PCs in traditional programs. We also include fixed effects for years of teaching experience at graduation to account for PCs who have prior teaching experience at time of graduation in all models that include experience.

We finally develop splines for the experience profile of each teacher. These splines group multiple years of experience and let us model growth in evaluation measures at different parts of the teaching career. The first part of the spline groups years 0 through 2 and the second part years 3 through 5. These two parts roughly overlap with the pre-tenure period and the early post-tenure period for teachers in our sample.

Program Features

We combine data from the state's program completer dataset with US 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 non-degree teacher certificate programs.

The state follows US department of education categorization of clinical placements as traditional student teaching as a field placement up to a semester in length, internship as a year-long field placement, and 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 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; this includes PCs in institutions that only offer job-embedded placements or institutions that also offer other kinds of placements. 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.

Finally, we use Title II data to measure TPPs' student teaching length, program size, and percentage of tenure track faculty. TPP leaders self-report these data yearly to the US Department of Education who then make them publicly available online. We divide each of these three variables into terciles and compare programs in the upper and middle tertile 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. For example, a teacher who would naturally receive high evaluations might choose to attend a specific TPP given its reputation. Similarly, preferences for certain preparation experiences could correlate in unexpected ways with TPP graduate employment decisions. Therefore, we

[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/

caution the reader not to interpret our results causally—they describe the association between attending a given TPP and later evaluation outcomes.

Analysis

RQ 1. Variation in initial effectiveness and growth across 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, multi-level 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 teacher 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. In an additional model, we include the years of experience spline to control for the effects of teaching experience on evaluation scores.

More formally, we estimate the following model

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

where τ_t and τ_p are the variance components explained at the teacher- and program-level respectively; σ^2 is the residual variance that is unaccounted by correlation between teacher- and program-level units.

We calculate the ICC values as the proportion of variance that is explained by intra-class correlations as

$$ICC_p = \frac{\tau_p}{\tau_t + \tau_p + \sigma^2}$$

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 effects on initial effectiveness and growth rates

We use a longitudinal, multilevel model to estimate the returns to attending a given TPP on teacher evaluation scores at different levels of experience. We note here that the next models for the next two research questions use a different random effects structure than the models for RQ1 because we aim to model the nesting of teacher observations within schools and districts in these models rather than the nesting of teachers in TPPs, like we did in RQ1.

Our preferred model is

$$Y_{isdt} = \beta_0 + \sum_{j=1}^{3} \beta_j \cdot f(exp)_{isdt} + \sum_{k=1}^{35} \gamma_k \cdot TPP_i + \sum_{l=1}^{68} \delta_l \cdot f(exp)_{isdt} \times TPP_i + \nu_d + \nu_s + \nu_i + \epsilon_{isdt}$$

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. 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 returns to attending a TPP to vary across experience levels.

γ_k and δ_l are the coefficients of interest. These coefficients are TPP fixed intercepts and slopes and estimate the returns to attending a TPP to vary across experience levels . γ_k capture differences among first-year performance for graduates from different TPPs. δ_l capture 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_j. Therefore, we report the coefficient $\gamma_k + \delta_l$ 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_s, and ν_i are nested random intercepts terms for each school district d, school s, and teacher i respectively. These terms capture the nested nature of the teaching profession 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 evaluation scores that is not accounted for by teacher preparation. 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 this experience profile in two ways. First, we use a piecewise linear spline to calculate the average growth in evaluation scores for given parts of the teaching career. This spline has three parts: 0-2, 3-5, and 6-9 years of experience. Using a spline allows us to calculate the returns to attending a TPP to vary across experience levels at the same time as allowing these return rates to vary among different parts of teachers' professional careers. This also allows us to assume that the returns to attending a TPP to vary across experience levels follow a linear functional form on only small parts of the experience profile. We 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 returns to attending a TPP to vary across experience levels, 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 Figure 4. We find that these two specifications give us virtually identical estimates for the returns to attending a TPP to vary across experience levels. We chose to use the linear spline specification for both interpretability and parsimony reasons.

RQ 3. TPP features that predict graduates' initial 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 who share common attributes (i.e., license type, degree level, clinical placement, student teaching length, program size, and faculty composition). We divide continuous variables

into tertiles when necessary to ensure consistency across our models. We first estimate the predictive power of these features in separate models for each feature; then we estimate a pooled model that includes all the indicator variables.

Bias versus Precision in Our Random Effects Models

We decided to use school random effects models over school fixed effects 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, let's take a school A that hires a few PCs every year and let's assume that these PCs have higher than average evaluation scores. A school fixed effects model averages out these PCs' high evaluation scores (i.e., difference out the group-level mean) and then leverages the remaining, within-school variation in PCs' evaluations. A random effects model balances the shrinkage of the school effects towards the population grand mean for schools with extreme fixed effects values or schools with few observations with the total variance in evaluation scores observed at each school. 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). School fixed effects models would absorb the TPP effects in both of these cases, leading to estimates that are over-corrected towards the state's grand mean.

The shrinkage in random effect models, however, could bias the estimates of the regression coefficients when an omitted variable correlated with the explanatory variables in the regression explains the extreme values in the school fixed effects that are shrunk. Coming back to our example, the principal at school A might, for instance, give higher evaluations to new teachers, regardless of their instructional performance. The random effects models will shrink the school effect towards the PC grand mean because school A is an extreme case with few observations. This could lead to bias estimates for the growth coefficients for this school, as we would only partially account for the principal's higher evaluation scores in the random effects models. On the other hand, school fixed effects models would average out this principal effect, thus effectively handling bias better in this case. To address this concern, we include a set of school-level covariates to control for effects of omitted variables onto our regression estimates. This allows us to control for the effects of omitted variables that are correlated to these school-level covariates. We believe that this set of covariates help reduce the risk of bias in our estimates while still retaining the better precision of random effects models. That said, unobserved characteristics may still bias our results.

As a robustness check, we present estimates for the total variance in evaluation scores that TPPs explain using both random effects models and two-step fixed effects models in RQ1. The two step models first calculate residualized outcomes from school and year fixed-effects models and then use these residuals as the outcome variables for the mixed effects regressions. We generally find that these two-step 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 step 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 returns at the tails of the distribution, we expect that this adjustment is not needed in our case and that the random effects models are an appropriate modeling approach for this paper.

Results

We report our results into three sections following our research questions. In research question 1, we find that TPPs explain about three percent of the variance in early career observation scores for TPP graduates and about one and a half percent of the variance in VAM scores. We expand on these initial findings in research question 2 by estimating TPP effects for initial effectiveness and later growth in evaluation scores. We find that differences between the top and bottom quintile TPPs is equal to about the growth we observe during the first two to three years of teaching. Finally, we explore how TPP features, such as student teaching duration or types of clinical placements, explain part of the variation we observe across TPPs. In the sections below, we provide more detailed answers to each of our research questions.

RQ 1. How much of the variance in early career teachers' initial effectiveness and growth is explained by teacher preparation programs?

We first report the results of the decomposition of the outcome variance from our multilevel models in Table 2. This table is organized by outcome. We report the results for observation ratings on the left and VAM scores on the right. The first two rows report the variance components for the two levels in our multilevel models—teachers and TPPs. We find that TPPs explain a non-zero portion of variance in our outcomes of interest. We find that 3.4 percent of the variance in observation ratings and 1.4 percent of the variance in VAM scores could be explained by differences across TPPs. These estimates are the upper bound of the outcome variance that could be explained by teacher-level and TPP-level random effects. In the third and fourth columns, we report the variance components after including the experience splines in the regression models. We observe that these variance components persist after introducing the experience splines. This suggests that TPP-level factors could capture time-invariant characteristics at each level that can explain differences in initial evaluation scores and that are not influenced by experience, as we find that the total variance explained in observation ratings changes from 3.4% to 3.2% and for VAM from 1.4% to 1.7% after we include the experience splines in the models. Finally, these models provide a more conservative estimate of the variance components that could be because of differences among TPPs after adjusting for graduates' experience levels.

We estimate the variance components using two different specifications for our outcomes of interest. The first set of estimates uses the standardized observation ratings and VAM scores. The second set of estimates uses the residualized observation ratings and VAM scores from models that control for school fixed effects. We notice that school-fixed effects account for about half of the variance that we observed across TPPs for observation ratings and account for most of the variance for VAM scores. This could suggest that TPP graduates tend to be employed in a similar set of 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.

Recent evidence suggests that the variance in evaluation outcomes explained by TPPs could be a result of estimation errors (von Hippel & Bellows, 2018). Our findings show that TPP-level differences still persist after accounting for evaluation scores differences among teachers and after accounting for school-level fixed effects. These results give us confidence in investigating the relationship between the returns of graduating from one TPP on graduates' initial performance and growth trajectories. We report these results in the next section and we discuss how estimation error does not explain all the differences we observe among TPPs in Tennessee.

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

We begin answering this question by considering the average state-wide growth trajectories for observation ratings and VAMs. These estimates are reported in Table 3. We report the results of four models: models (1) and (2) report estimates for observation ratings, models (3) and (4) report the estimates for VAM scores. The null models estimate the average evaluation scores for TPP graduates and provide unconditional estimates of the variance components in our multilevel models. The growth curve models report the estimates for the first year of teaching average evaluation scores and year-by-year early career growth in evaluation scores.

We find that the average growth rate across TPPs is higher early in graduates' careers for both outcomes, and that this growth rate tapers off with time. This result is in line with what Papay and Kraft (2015) reported at the individual teacher level. Interestingly, it takes teachers about five years of experience to reach the average state-wide observation ratings, but only three years of experience to reach the average state-wide VAMs. We now turn to our research questions and explore differences in initial effectiveness and growth rates among teacher preparation programs.

Figure 1 displays the variation in initial performance by TPP. We find that initial observation ratings are spread over about 0.6 standard deviations and that VAMs are spread over about 0.3 standard deviations. These correspond with returns to two years of teaching experience for observation ratings and about three years of experience for VAMs. Said another way, we observe that there is significant variation in the initial effectiveness of graduates from different TPPs and that these differences could be equal to several years of teaching experience.

Von Hippel and Bellows (2018) have argued that estimates for the effects of TPPs on VAMs could be just due to random measurement error. They propose a test to measure the extent to which TPP effects follow the expected null distribution under the assumption of random measurement errors. We report the results of these tests in Figure 2. Each panel reports the results of a test for initial effectiveness or growth estimates for observation ratings and VAMs. Each figure displays the point estimate for each TPP as a black diamond, 95% confidence intervals are displayed as the solid whisker around the point estimate, and the Bonferroni adjusted 95% confidence intervals are reported by the dashed whisker and the gray dots. We report the results of each test below the main plot. These statistics include 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 that could be explained by differences across TPPs instead of estimation error.

We find that the p-values for all the six tests reject the null hypotheses that the distribution of our estimates for TPP initial effectiveness follows the null distribution under random estimation error. This means that the differences in initial effectiveness and early career growth rates among TPP graduates are not a result of estimation error. We also find that our estimates for initial observation ratings seem to be more reliable in distinguishing TPPs, $\rho = 0.79$, than our estimates for VAMs, $\rho = 0.70$. Taken together, these results give us confidence that our estimates are able to identify real differences among TPPs both in terms of initial levels and subsequent growth instead of these differences being the result of random measurement error associated with the outcomes of interest.

Figure 3 displays the growth trajectories for graduates from four selected TPPs in the state. Focusing on the growth trajectories for observation ratings, we find two results to be noteworthy. First, we observe that TPPs that start at similar levels of effectiveness have different growth trajectories. For example, the yellow and red TPPs start at similar levels of initial effectiveness. However, graduates from the yellow TPP grow at faster rates (on both outcomes) than graduates from the red program. Comparing across outcomes, we note that the graduates from the green TPP have higher observation ratings and VAMs than other graduates. However, their growth trajectories appear to be different. The observation ratings for these graduates continue to grow, outpacing the rest of the teachers in the state. For VAMs, we observe that these graduates experience a rapid decline in later parts of their careers. While this rapid decline could be due to attrition, this underscores the fact that growth trajectories capture important aspects of teacher preparation that might not be considered by average effectiveness measures.

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, we find initial effectiveness to be positively and moderately correlated ($r = 0.28$) with the growth rate during the first three years of teaching. This suggests that teachers that receive initially higher observation ratings tend to also grow at higher rates early in their careers, but these correlations are non-significant. Also, this positive correlation gives suggestive evidence that regression to the mean in observation ratings might not happen during this time. In the presence of regression to the mean, we would expect teachers who receive lower than the mean scores would grow at faster rates and teachers who receive higher than the mean scores would grow at slower rates, leading to a negative correlation between initial effectiveness and growth rates.

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 negative and moderately 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, an observation that could partially be explained by the regression to the mean argument for teacher evaluation scores (Atteberry et al., 2015).

The correlation patterns for VAMs are somewhat different. We find that initial performance 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 ($r = -0.70$).[3]

Finally, we note that growth early in the career is negatively correlated with growth later in the career for observation ratings ($r = -0.211$) and for VAMs ($r = -0.26$). This result is in line with our prior results suggesting that the growth rates taper off with experience.

RQ 3: What program-level features predict graduates' initial effectiveness and growth during their early teaching careers?

Figure 5 and Appendix Table 1 report the results of the relationship between different TPP features and estimates for graduates' initial effectiveness and growth on observation ratings and VAMs. We estimate these coefficients by including variables representing salient features of TPPs in our preferred models. We report estimates for six features: certification area, program type, clinical placement, student teaching length, program size, and percent of tenure track faculty. In this section, we focus our discussion on estimates from models focused on first-year levels of effectiveness and initial (years 0-2) growth trajectories; see Appendix Table 1 for estimates from models using later (years 3-5) growth trajectories. Finally, since estimates from separate and combined models are similar (see Appendix Table 1), we focus here on estimates from separate models.

Observation Ratings

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, while multiple and "other" credential teachers receive higher initial ratings. However, elementary teachers exhibit greater subsequent growth when compared to teachers in all other endorsement areas.

[3] It is worth noting here that the results that we observe in Figure 2 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 performance are carried over during their early career period. Were regression to the mean to exist, we would expect (1) there to be significant differences between TPPs during early growth (years 0-2) and (2) 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' VAM scores 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 non-significant 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.

We find no differences in initial observation ratings between teachers who complete post-baccalaureate programs and teachers who complete four-year, baccalaureate programs. In addition, we find no differences in subsequent growth trajectories between these groups.

Looking at clinical placement structure, PSTs who complete internship and job-embedded placements initially outperform PSTs who complete traditional student teaching placement. Subsequently, PSTs who complete internship placements also demonstrate faster rates of growth while PSTs who complete job-embedded demonstrate slower growth rates.

Graduates from TPPs that have longer clinical placements (i.e., in the upper third of duration) tend to receive higher initial observation ratings than graduates who complete shorter placements. Though they have initially stronger observation ratings, graduates who complete longer placements have significantly slower rates of subsequent growth.

Graduates from larger, and to a lesser degree middle-sized, programs have lower levels of initial effectiveness but subsequently greater rates of growth. Results related to the share of tenure track faculty are more mixed. Compared to graduates from programs with small shares of tenure track faculty, those from programs with large shares have stronger initial effectiveness but lower growth rates in subsequent years. Graduates from programs with a medium share of tenure track faculty have initially lower effectiveness but have comparable rates of subsequent growth.

VAMs

Turning to VAM scores rather than s observation ratings, graduates with secondary (non-STEM) endorsements have higher levels of initial effectiveness when compared to graduates with elementary endorsements, while graduates with special education endorsements have initially lower levels. Graduates from all non-elementary endorsements have statistically similar rates of initial growth as graduates with elementary endorsements except for secondary STEM teachers who grow at faster rates.

Compared to graduates from undergraduate programs, graduates from post-baccalaureate programs, whether degree conferring or not, have similar levels of initial effectiveness; graduates from post-baccalaureate (no degree) programs, though, have subsequently lower growth rates.

Regarding placement type, graduates who completed internship and job-embedded placements have greater initial levels of effectiveness than graduates who complete traditional, student teaching placements. However, the subsequent growth rates of all groups are statistically similar. Notably, graduates who completed traditional student teaching placements have significantly greater growth rates than other graduates later on (between three and five years of teaching).

Duration of clinical placement is unrelated to initial effectiveness. However, graduates who completed the longest clinical placements tend to have slower, subsequent growth rates.

Both program size and the percentage of tenure-track faculty are unrelated to initial levels of effectiveness as well as subsequent growth rates.

Discussion

We find significant between-TPP variation in graduates' initial levels of instructional effectiveness in both classroom observation ratings and teacher VAM scores. We also find significant and meaningful differences between TPPs in terms of graduates' growth in instructional performance as measured by both observation ratings and VAMs and subsequent growth during their first five years of teaching experience. These results regarding both levels and growth are also robust to the von Hippel and Bellows (2018) Q-statistic test, further indicating that our observed variation is likely due to *real* differences between TPPs rather than measurement error. Moreover, we find that the initial effectiveness estimates and subsequent growth estimates are only weakly correlated—in other words, programs that seem to promote initial effectiveness are not necessarily those where teachers improve more early in their careers.

On the other hand, our results also show that only a small portion of total variance in TTP graduates' evaluation outcomes is explained by differences between TPPs. We estimated that about 3.4% of the variance in initial observation ratings and 1.4% of the variance in initial VAM scores is due to different returns to TPPs. These estimates are similar in magnitude to the amount of variance that our models explain between school districts in the state and to what others have reported for the returns of 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 re-analyses have suggested differences between TPPs observed in some prior studies may be due to noise, our Q-statistic tests suggest differences to be real (signal). However, whether this amount in total variance explained is large enough to be policy relevant still remains unclear.

Regardless of the absolute size of explained variance, the difference in the returns of TPPs for teachers who graduate from programs in the top decile and the statewide average is about 0.15 standard deviation units for both observation ratings and VAM scores. This difference is roughly equal to about half the growth that new teachers experience in one school year. Compared to graduating from an average TPP, we find that graduating from the best TPPs in the state is comparable to about four and a half months of teaching experience.

Thus, some TPPs seem to graduate teachers who are initially more instructionally effective than other TPPs. At the same time, other TPPs also graduate teachers who, while not necessarily instructionally effective initially, are able to improve at faster rates than graduates from other TPPs. This might suggest that some programs specialize in preparing teachers to perform effectively out of the gate, while others prepare teachers to learn more effectively on the job.[4]

[4] 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.

It also suggests that policymakers should take into consideration both of these factors—initial preparation and later growth—rather than focusing on average initial performance 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, efforts to evaluate TPPs based on the effectiveness of their graduates should factor in not only initial effectiveness but also growth over time and retention 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 the first few years of the career may obscure programs that set teachers up for longer-term success. A more nuanced evaluation policy would likely provide more comprehensive feedback to TPPs about graduates' initial performance as well as growth over time so that program leaders have more information about how to improve.

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 driving initial effectiveness while others may be driving growth, or whether there may be some features that predict both initial effectiveness and growth. Consistent with past work (e.g., Ronfeldt et al, 2020), none of the TPP features we studied were consistently related to the four main outcomes in our analyses—observation rating levels, observation rating growth, VAM levels, VAM growth.

Before elaborating on which TPP features predict levels versus growth in teaching effectiveness (see below), it is important to acknowledge a key limitation of this study—that various forms of selection may explain results across research questions and analyses. Forms of selection likely at work include selection of prospective teachers to TPPs (and, thus, TPP features), 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. For example, it is possible that more promising candidates are recruited by or select into certain TPPs; thus, finding graduates from these TPPs to perform better may be due to selection rather than impacts of the TPPS on graduate performance. Likewise, where we find certain TPP features (below) to be associated with graduates' instructional effectiveness, we cannot necessarily infer that these associations represent causal relationships given that more promising candidates may have initially selected into TPPs with these features. Despite this, we believe it is still important to establish correlational evidence between TPP features and outcomes to guide future experimental or causal inference studies in establishing these correlations to be causal in nature (or not).

Since we are interested in whether some features predict initial effectiveness, while others predict subsequent growth, below we begin our discussion of features by considering features that predicted initial effectiveness on both observation ratings and VAMs. After we consider those features that predict subsequent growth on both measures.

Features Related to Initial Performance

There were three features that were associated with initial effectiveness on both observation ratings and VAMs: endorsement area, job-embedded placements, and internship placements. Compared to graduates with elementary endorsements, graduates endorsed in special education had lower initial observation ratings, lower growth on observation ratings, and lower initial VAMs (Bardelli & Ronfeldt, 2020). However, 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 (Campbell & Ronfeldt, 2018; Buzick & Jones, 2015; Jones & Brownell, 2014). Even though we control for school characteristics in our models, we are unable to adjust for classroom characteristics (e.g., having a high proportion of students who qualify for special education services). Thus, these results could be driven, at least in part, by unobserved classroom characteristics. More research is needed to examine this.

The other two features involved the type of clinical placement. Specifically, graduates who completed internship or job-embedded placements had significantly better initial performance in both observation ratings and VAMs as compared to graduates who had completed traditional student teaching placements. Again, we caution against causal interpretations. Since job-embedded placements are often associated with alternative route programs and internship (year-long) placements with residency programs, it is especially important to emphasize that these results cannot yield causal conclusions about these pathways. Each type of program includes many other program features that could explain observed effects of clinical placement type, including that both residency and alternative route programs tend to provide ongoing mentoring after graduates become inservice teachers of record while traditional programs tend not to. Moreover, it is important to acknowledge that there is much variation in outcomes among programs that utilize job-embedded placements; likewise for programs that utilize internship and student teaching placements.

Features Related to Early Growth (Years 0-2)

Only one program feature predicted early growth on both observation ratings and VAMs - clinical placement duration. Specifically, compared with graduates that completed the shortest (fewest weeks) clinical placements, graduates that completed the longest had significantly lower rates of subsequent growth. One possibility is that the latter group had already learned so much during their longer clinical placement that they had less room for growth. Consistent with this explanation, graduates who completed the longest placements had significantly greater initial observation ratings than graduates who completed the shortest placements.

Student teachers who completed an internship continued to grow at a faster rate than student teachers that completed a traditional student teaching placement. This result is in line with prior research that has found that the quality of student teaching experiences, rather than their duration, is associated with increased teachers' perceptions of instructional preparedness, efficacy, and career plans (Ronfeldt & Reininger, 2012).

Features Related to Later Growth (Years 3-5)

Though we focus less in our results section on later (years 3-5) growth, we include a discussion here of features predicting later growth since it has implications for clinical placement type, which we find to be associated with initial effectiveness. Specifically, while we find graduates that completed traditional student teaching placements to have lower initial effectiveness than graduates who completed job-embedded and internship placements (see above), they have significantly greater later (years 3-5) growth on both observation ratings and VAMs.

Why graduates in this group seem to be less effective early on but grow at substantially greater rates later in their careers is beyond the scope of the present study and an area in need of future research. That said, we propose two 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, Loeb, & Wyckoff, 2015). Related, this could suggest that candidates from traditional programs 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, 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).

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

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

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

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

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*. Long Beach, CA.

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

Council of chief state school officers (CCSSO). (2019). *Accountability in teacher preparation: Policies and data in the 50 states & DC*. https://ccsso.org/sites/default/files/2017-10/50StateScan092216.pdf

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*. The University of Chicago Press.

Feng, L., & Sass, T. R. (2013). What makes special-education teachers special? Teacher training and achievement of students with disabilities. *Economics of Education Review*, *36*, 122–134. https://doi.org/10.1016/j.econedurev.2013.06.006

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

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 re-examination 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

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

Master, B., Loeb, S., Whitney, C., & Wyckoff, J. (2016). Different skills?: Identifying differentially effective teachers of English language learners. *The Elementary School Journal*, *117*(2), 261–284. https://doi.org/10.1086/688871

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. *J. Public Econ.*, *130*, 105–119. https://doi.org/10.1016/j.jpubeco.2015.02.008

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

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., & 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.

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. In *UNC Charlotte Economics Working Paper Series*. https://belkcollegeofbusiness.uncc.edu/economic-working-papers/wp-content/uploads/sites/850/2018/06/wp2018-001.pdf

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)
VAM Scores	-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

Notes. This table reports descriptive statistics for the teachers in our sample. The outcomes of interests are standardized using the state-wide 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.

Table 2: Variance Decomposition

		Observation Ratings		Value-Added Measures	
		Raw	**Two-Step**	**Raw**	**Two-Step**
Null model	TPP	0.034	0.015	0.014	0.004
	Teacher << TPP	0.643	0.578	0.381	0.312
+ Experience spline	TPP	0.032	0.017	0.017	0.006
	Teacher << TPP	0.643	0.540	0.380	0.307

Note. Null models do not include any covariates. Models with the experience spline include controls for years of experience. 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.

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

	Observation Ratings		Value-Added Measures	
	(1)	(2)	(3)	(4)
	Null Model	**Growth Curve**	**Null Model**	**Growth Curve**
Average Evaluations	-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	123553	123553	47491	47487
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.

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: Correlations between Intercepts and Slope Terms		
	Initial Effectiveness	**Growth 0-2 Years of Experience**
Observation Ratings		
Growth 0-2 Years of Experience	0.2826	-
Growth 3-5 Years of Experience	-0.2778	0.2047
Value-Added Measures		
Growth 0-2 Years of Experience	0.1019	-
Growth 3-5 Years of Experience	-0.7040*	-0.2568

Figure 1: Differences in Initial Performance by TPP

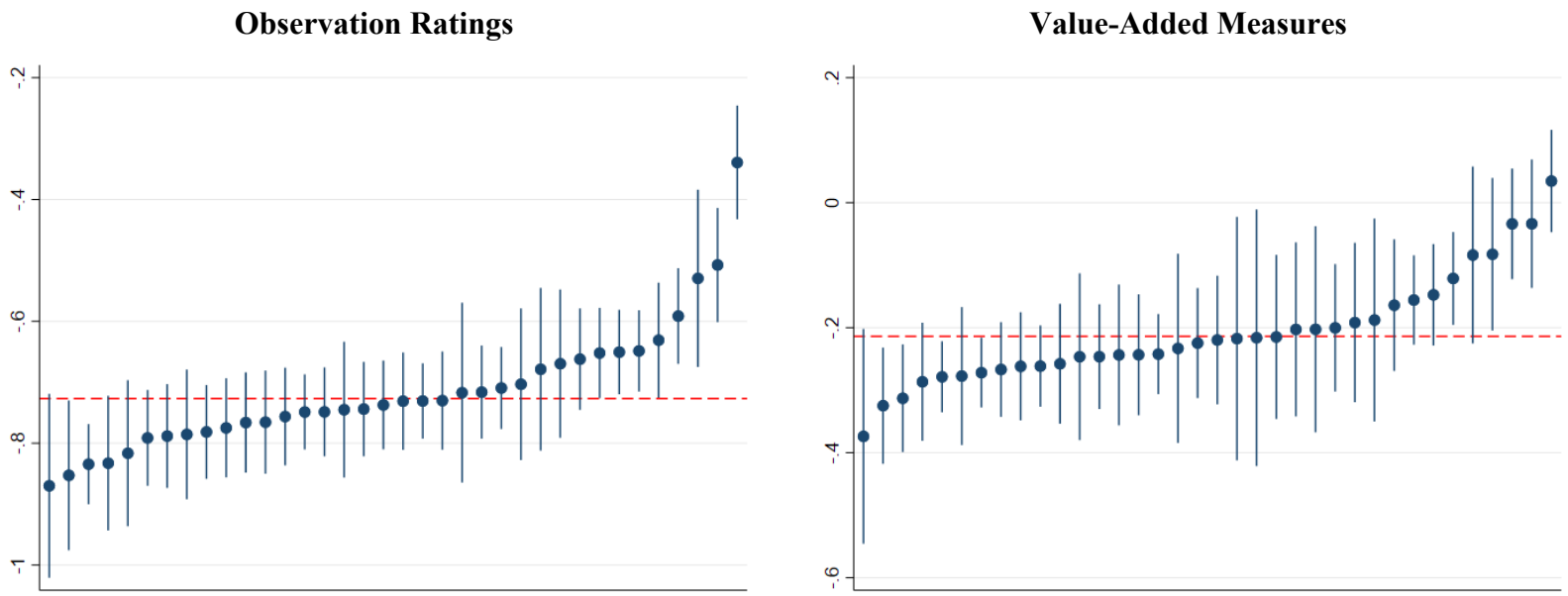


Figure 2: Test of Non-random Distribution of the TPP Effects

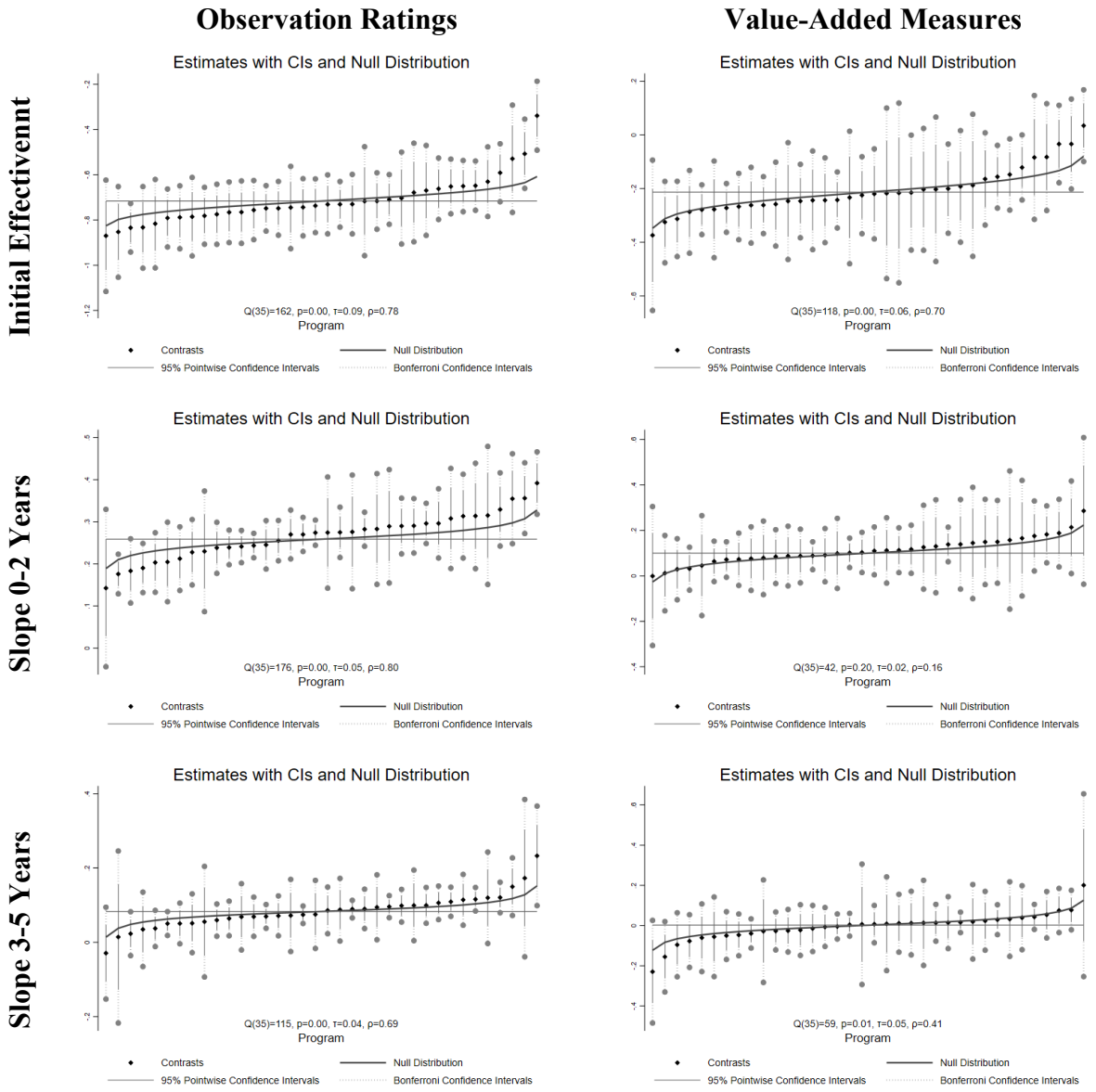


Figure 3: Difference in Observation Ratings Growth Trajectories for Four Selected TPPs

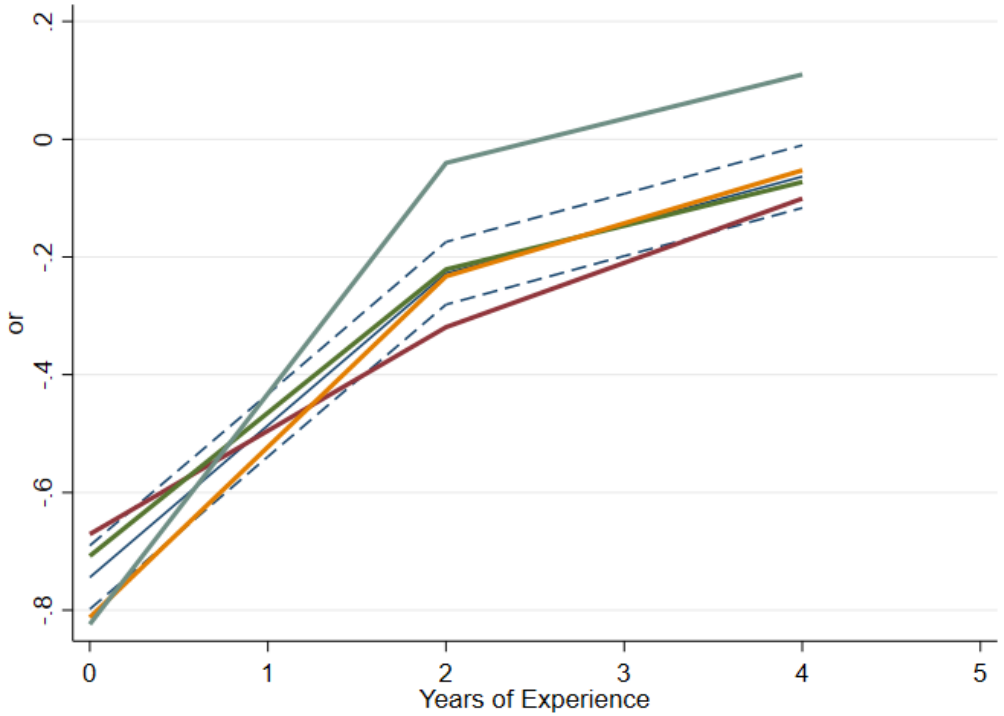


Figure 4: Comparison between Spline and Indicators Experience Models

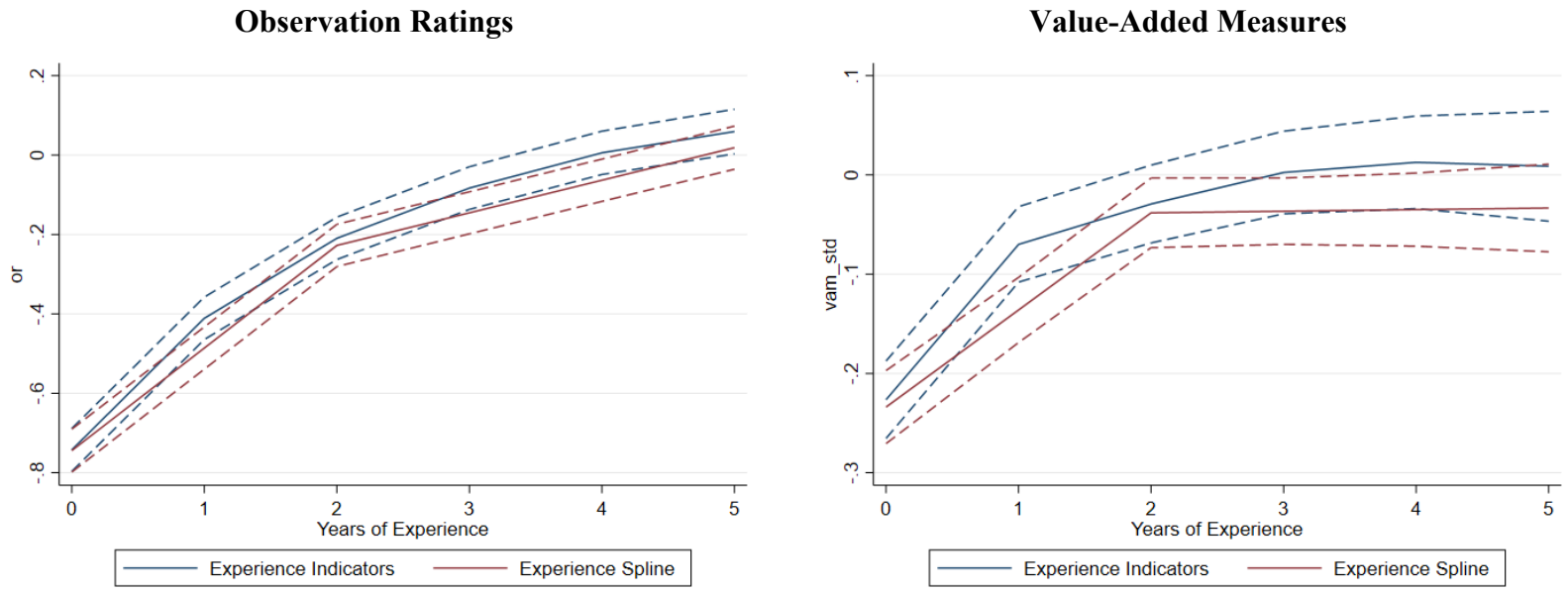
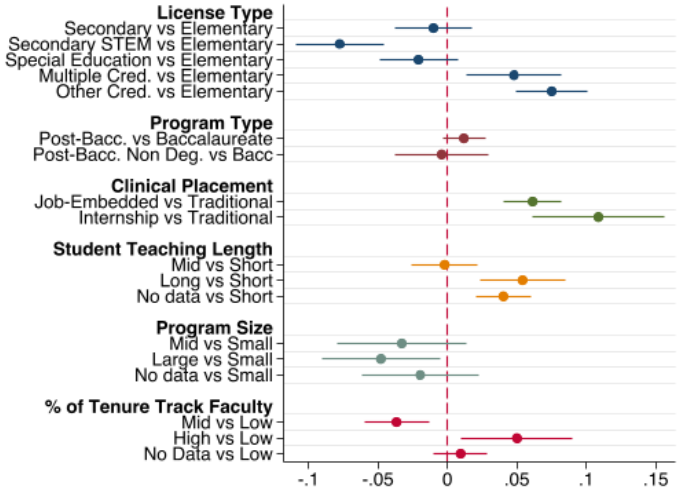


Figure 5: Relationship between Features of TPPs and Growth Estimates

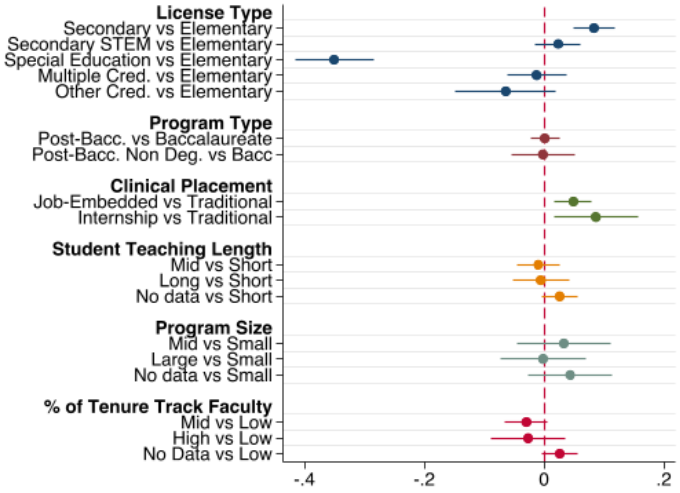
Observation Ratings

Initial Effectiveness

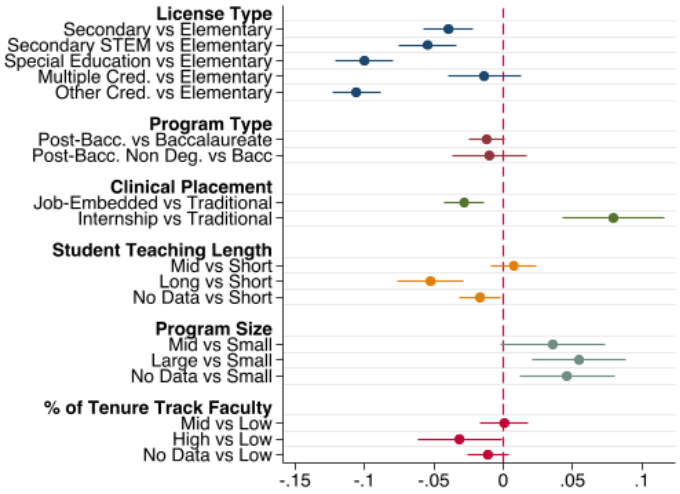


Value-Added Measures

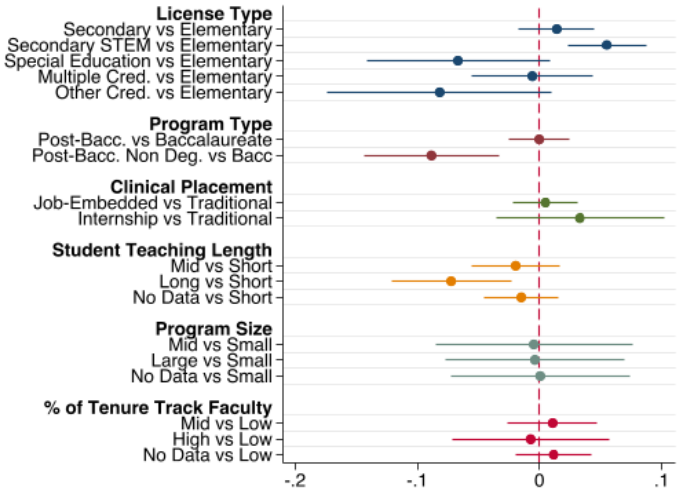
Initial Effectiveness



Growth Estimate 0-2 Years



Growth Estimate 0-2 Years



Appendix Table 1 – Relationship between Features of Teacher Preparation Programs and Initial Effectiveness and Slope

	Initial Effectiveness				Growth 0-2 Years				Growth 3-5 Years			
	Observation Ratings		Value-Added Measures		Observation Ratings		Value-Added Measures		Observation Ratings		Value-Added Measures	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(5)	(6)	(7)	(8)
	Separate Models	Combined Models	Separate Models	Combined Models	Separate Models	Combined Models	Separate Models	Combined Models	Separate Models	Combined Models	Separate Models	Combined Models
Credential (vs Elementary)												
Secondary	-0.007	-0.011	0.085***	0.082***	-0.042***	-0.039***	0.018	0.021	-0.006	0.003	-0.054***	-0.049***
	(0.015)	(0.015)	(0.018)	(0.018)	(0.009)	(0.009)	(0.016)	(0.016)	(0.007)	(0.007)	(0.014)	(0.014)
Secondary STEM	-0.073***	-0.094***	0.026	0.010	-0.057***	-0.050***	0.057***	0.060***	0.002	0.020*	-0.027+	-0.009
	(0.017)	(0.017)	(0.020)	(0.020)	(0.011)	(0.011)	(0.017)	(0.017)	(0.008)	(0.009)	(0.014)	(0.015)
Special Education	-0.026+	-0.038**	-0.367***	-0.381***	-0.099***	-0.091***	-0.045	-0.045	-0.039***	-0.024**	0.047	0.065*
	(0.015)	(0.015)	(0.034)	(0.034)	(0.011)	(0.011)	(0.039)	(0.039)	(0.009)	(0.009)	(0.030)	(0.031)
Multiple Credentials	0.048**	0.044*	-0.008	-0.024	-0.020	-0.009	-0.004	0.000	0.017	0.017	-0.036	-0.019
	(0.018)	(0.018)	(0.026)	(0.026)	(0.014)	(0.014)	(0.026)	(0.027)	(0.012)	(0.012)	(0.028)	(0.029)
Other Credentials	0.069***	0.069***	-0.070	-0.064	-0.104***	-0.099***	-0.081+	-0.082+	-0.025***	-0.018*	0.023	0.035
	(0.014)	(0.014)	(0.043)	(0.043)	(0.009)	(0.009)	(0.047)	(0.047)	(0.007)	(0.007)	(0.044)	(0.044)
Program Type (vs Bachelor's)												
Post-Baccalaureate Degree	0.013	-0.025**	0.004	-0.021	-0.016*	0.014	-0.005	-0.004	-0.026***	-0.007	-0.027*	-0.003
	(0.008)	(0.009)	(0.013)	(0.014)	(0.007)	(0.009)	(0.013)	(0.016)	(0.005)	(0.006)	(0.011)	(0.013)
Post-Baccalaureate No Degree	-0.006	-0.069***	-0.003	-0.048+	-0.011	0.041**	-0.085**	-0.098**	0.004	0.041+	0.047	0.097*
	(0.017)	(0.019)	(0.027)	(0.029)	(0.013)	(0.015)	(0.028)	(0.031)	(0.020)	(0.021)	(0.046)	(0.047)
Clinical Placement (vs. Traditional)												
Internship	0.125***	0.120***	0.102**	0.082*	0.075***	0.068***	0.019	0.017	0.004	0.017	-0.174**	-0.160**
	(0.026)	(0.027)	(0.038)	(0.039)	(0.019)	(0.020)	(0.036)	(0.038)	(0.039)	(0.039)	(0.060)	(0.061)
Job Embedded	0.066***	0.076***	0.054***	0.075***	-0.036***	-0.007	-0.001	0.017	-0.043***	-0.040***	-0.043***	-0.046**
	(0.011)	(0.013)	(0.016)	(0.018)	(0.007)	(0.009)	(0.014)	(0.017)	(0.006)	(0.007)	(0.013)	(0.015)
Length of Student Teaching (vs. Quartile 1)												
Quartile 2	0.002	0.008	-0.011	-0.017	0.009	0.008	-0.016	-0.021	0.004	-0.001	0.010	0.013
	(0.012)	(0.013)	(0.019)	(0.019)	(0.008)	(0.009)	(0.018)	(0.019)	(0.007)	(0.007)	(0.016)	(0.016)
Quartile 3	0.058***	0.050**	-0.007	-0.026	-0.051***	-0.044**	-0.067**	-0.068*	-0.005	0.016	-0.019	-0.016
	(0.016)	(0.017)	(0.024)	(0.026)	(0.012)	(0.014)	(0.025)	(0.027)	(0.010)	(0.011)	(0.023)	(0.025)
No Data	0.052***	0.057**	0.027+	-0.053+	-0.021**	-0.034+	-0.015	-0.010	-0.025***	-0.001	-0.024+	0.004
	(0.010)	(0.021)	(0.016)	(0.030)	(0.008)	(0.017)	(0.016)	(0.032)	(0.006)	(0.017)	(0.014)	(0.037)
Program Size (vs. Quartile 1)												
Quartile 2	-0.031	-0.010	0.034	0.027	0.034+	0.018	-0.008	0.004	-0.008	-0.023	0.055	0.050
	(0.024)	(0.025)	(0.040)	(0.041)	(0.019)	(0.020)	(0.041)	(0.043)	(0.017)	(0.018)	(0.038)	(0.040)
Quartile 3	-0.049*	-0.012	-0.002	-0.011	0.054**	0.015	-0.007	-0.009	0.016	-0.004	0.049	0.033
	(0.022)	(0.024)	(0.037)	(0.038)	(0.017)	(0.019)	(0.038)	(0.040)	(0.015)	(0.017)	(0.035)	(0.038)
No Data	-0.013	-0.001	0.048	0.049	0.039*	0.044*	-0.009	-0.019	-0.012	-0.022	0.022	0.024
	(0.022)	(0.026)	(0.036)	(0.043)	(0.018)	(0.022)	(0.038)	(0.045)	(0.015)	(0.019)	(0.035)	(0.045)

Percent of Tenure-Track Faculty (vs. Quartile 1)												
Quartile 2	-0.038**	-0.030*	-0.033+	-0.039*	-0.000	-0.007	0.011	0.002	0.010	0.011	-0.012	-0.021
	(0.012)	(0.012)	(0.018)	(0.019)	(0.009)	(0.009)	(0.019)	(0.019)	(0.007)	(0.008)	(0.017)	(0.017)
Quartile 3	0.047*	0.041+	-0.029	-0.037	-0.032*	-0.023	-0.006	-0.020	-0.010	0.003	0.002	0.006
	(0.020)	(0.022)	(0.031)	(0.033)	(0.015)	(0.016)	(0.033)	(0.035)	(0.012)	(0.013)	(0.027)	(0.029)
No Data	0.011	-0.032*	0.028+	0.007	-0.017*	-0.017	0.008	-0.000	-0.014*	0.014	-0.030*	-0.011
	(0.010)	(0.015)	(0.015)	(0.024)	(0.008)	(0.011)	(0.016)	(0.025)	(0.006)	(0.010)	(0.014)	(0.022)
Covariates	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Experience Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	97785	96459	44599	44309	97785	96459	44599	44309	97785	96459	44599	44309

Note. Standard errors in parentheses. + p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001