Math and Science Outcomes for Students of Teachers from Standard and Alternative Pathways in Texas

November 29, 2016

Michael Marder[1] and Caitlin Hamrock[2]
[1]Department of Physics and UTeach, The University of Texas at Austin, 78712, marder@chaos.utexas.edu
[2]Department of Sociology, The University of Texas at Austin

Abstract

We assess the impact of teachers from different preparation pathways on Algebra I and Biology learning outcomes in Texas. Data come from the state of Texas for academic years 2010-2011 and 2011–2012. We examine both novice and experienced teachers. We make three sets of comparisons, ranging from broad pathways to specific programs. First we compare teachers from all standard university programs to all teachers from alternative certification programs. Second, we select teachers from a collection of leading universities and compare them with alternatively certified novice teachers for the same subjects in the same schools. Third, we repeat this process for teachers from UTeach, a STEM-specific university-based program with 8 sites in Texas. Students whose teachers came from a standard pathway gain around one more Month of Schooling than those whose teachers followed an alternative pathway. Effects are larger for mathematics than for science. For some subgroups including Economically Disadvantaged and Gifted students, the advantage of having teachers from standard programs may be as large as 6 to 9 Months of Schooling.

Introduction

Subject	***Number of Teachers***	***Percent with no major in main assignment or not certified***
Mathematics	144,800	38%
Science	126,300	27%
Biology	51,900	35%
Physical Science	64,600	62%
Chemistry	24,300	66%
Earth Sciences	12,400	68%
Physics	13,300	63%

Table 1: STEM Teachers out of field in main assignment or not certified. Source: Schools and Staffing Survey (2012)

Persistent teacher shortages (Barth et al., 2016) make it difficult to offer high-quality mathematics and science courses to all high school students in the United States. One indication of the extent of national teacher shortages in STEM fields appears in Table 1. Nearly 40% of mathematics teachers either lack full teaching certification or lack a major or minor in mathematics. In the physical sciences, over 60% of teachers lack one or the other of these qualifications. Estimating from students taking Advanced Placement Computer Science (CS) exams (College Board, 2016), less than 20% of US high schools even offer computer science, and CS teacher shortages are difficult to monitor from Federal statistics because the subject is lumped in with mathematics. The shortage of STEM teachers may become greater because the number of teachers prepared in the highest-producing states has been falling rapidly (Figure 1).

Broadly speaking there are two main ways to address teacher shortages: retain existing teachers in the profession or produce new ones. To produce new teachers there are again two broad approaches. The first is to increase support to individuals entering teaching and to the traditional institutions that prepare them. The second is to reduce barriers to entering the teaching profession, and promote new pathways to teaching, including market-based solutions.

Why should new pathways be considered? The traditional source of newly credentialed teachers is public colleges and universities, and the quality of programs at these institutions has come into question (Greenberg et al., 2013). It cannot be taken for granted that increasing the number of new teachers from this pathway is desirable. Even if it is desirable, there is no agreed way to accomplish it. The largest program at the Federal level to increase the number of STEM teachers from universities is NSF's Noyce Scholarship program (National Science Foundation, 2016). Annual funding is around $56 million. The most important element of the program is scholarships and stipends given to current or former science and mathematics majors who commit to teach in high-needs districts. Award amounts vary, and not all the funds end up in scholarships, but estimating $15,000 per student puts an upper bound of 3700 teachers per year; actual numbers produced are probably below this value. According to the Schools and Staffing Survey (Table 1) around 180,000 STEM teachers are underqualified for their positions, meaning that the nation's largest scholarship program supports a number of teachers per year corresponding to less than 2% of a lower bound on national need. Thus the institutions that have traditionally supplied the United States with teachers are not succeeding in supplying enough teachers in STEM shortage areas, while programs intended to rectify these problems are doing so on too small a scale to solve the problem.

Such problems have persisted for decades (National Commission on Excellence in Education, 1983), which has led to calls to provide teachers in new ways (Hess, 2001). Alternative pathways to teaching now exist in all states, but there is great variation in the regulations that control what they are able or not able to do. In the Every Student Succeeds Act (114th Congress, 2015), Alternative Certification is mentioned over 30 times, with funding streams aimed at strengthening it, indicating that it is likely to become increasingly important.

Since the United States may be at the edge of pressing for a new expansion of alternative certification pathways, including those operating for profit (Teachers of Tomorrow, 2016), it is valuable to examine available evidence on the advantages and disadvantages of going in this direction. In contrast to reports that paint a bleak picture of traditional teacher certification (Greenberg et al., 2013; Duncan, 2010), and in contrast to an equally negative perception that many university faculty hold of alternative certification providers (Kamnetz, 2014), the findings from the research literature are moderate and mixed. Overall, teachers prepared through alternative certification pathways are less likely to remain in the teaching profession in their early years than those coming through standard routes. Some studies find a moderate advantage for students whose teachers came from traditional pathways, some find a moderate advantage for students whose teachers came through alternative routes, and some are unable to discern a difference. We will review these findings in the next section.

Before stating our research question, it will be helpful to provide more specific information on policy changes in progress that will affect the supply of new STEM teachers across the nation. The US Department of Education has directed every state to develop ratings of each Teacher Preparation Program (TPP). States must "make meaningful differentiations in teacher preparation program performance using at least three performance levels — low-performing teacher preparation program, at-risk teacher preparation program, and effective teacher preparation program (US Department of Education, 2016a, p. 670). There is a number of mandatory indicators to consider. The first of them is student learning outcomes. "For each year and each teacher preparation program in the State, a State must calculate the aggregate student learning outcomes of all students taught by novice teachers." The definition of novice teacher is "A teacher of record in the first three years of teaching who teaches elementary or secondary public school students" (US Department of Education, 2016a, p. 656). In addition, the State must consider employment outcomes, including placement and retention in high-needs schools, and employer and novice teacher satisfaction. In this paper we primarily provide information on the relationship between teacher preparation pathways and student learning outcomes, but we also provide some information concerning retention.

The rating of an individual TPP will have consequences. A program rated low-performing will lose access to federal funds, including TEACH grants for its students. A more serious consequence would be the loss of authority granted by each state to recommend students for certification. Such a decision is not discussed in the federal guidance, but is an obvious possibility.

In addition to the forthcoming ranking of every TPP, states frequently issue rules that govern the operation of teacher preparation pathways. For example, in October 2016 Louisiana adopted a rule change that extends student teaching for university-based programs to a full year and reduces the requirements for prior coursework (Sentell, 2016a). One possible outcome is that "the new rules will push more students into alternative certification programs, which already account for about half of the state's teachers" (Sentell, 2016b). The point here is not to take a position on residency programs, but simply to note that state regulation has the power to steer prospective teachers to enter one type of preparation pathway or another, most obviously but not exclusively through the way it governs alternative certification.

It is also possible for states to take supportive action towards standard teacher preparation programs. For example, UTeach is a STEM teacher preparation program that began at The University of Texas at Austin in 1997, and has expanded to 45 universities across the country. The expansion has been made possible by a variety of public-private partnerships, but in Texas, Arkansas, Florida, Tennessee, California, Massachusetts, and Georgia, it was partly due to a coordinated state effort that included state and federal funding and regulatory assistance. We will specifically study UTeach in Texas in this paper.

Our main research question is:

- What is the effect of teachers from different pathways on high school student learning outcomes in math and science? In particular, how do learning outcomes depend upon whether the teacher came through an alternative certification program or a standard university-based program?

The setting for our study is Texas, which as shown in Figure 1 has been producing more teachers than any other state. Texas is large and varied, making it possible to access a wide range of school environments — including urban, suburban, small town, and charter schools — a wide degree of variation in student socioeconomic status, and a large and varied collection of teacher preparation programs. Indeed Texas presents a unique opportunity to study alternative teacher certification pathways not involving universities because as shown in Figure 1 no other state approaches Texas in the number or proportion of teachers coming from these routes.

In formulating our approach, we were influenced by previous studies in Texas that encountered great difficulty in finding significant differences in student learning outcomes they could attribute to specific programs (Mellor et al., 2008; von Hippel et al., 2014). There is a tension between focusing upon specific programs, in which case sample sizes can be too low to provide enough statistical power to detect effects, and creating very broad categories, in which case essentially dissimilar programs may be grouped together. We present three program groupings: a large scale where all university-based programs are grouped together, a medium scale that groups together leading universities from Texas' flagship university systems, and a smaller-scale grouping of the teachers coming from Texas UTeach programs. We also made use of two levels of teacher experience, looking first at novice teachers — those with less than four years of experience — and then at experienced teachers, defined as those with up to ten years of experience. These levels of experience have specific connections to policy considerations and the history of teacher preparation in Texas, but also allow us to move back and forth between considering a larger group of teachers with much internal variation, and a smaller group with less variation, but with smaller samples and more uncertainty. The groupings of programs and years of experience let us examine student learning outcomes as if with a microscope that focuses in and out at various levels of magnification, searching for meaningful signals from a complex and noisy system.

Background

Value-Added Modeling

There is an extensive literature on value-added modeling of student learning, on alternative certification pathways, and of the use of value-added models to investigate pathways. Early work on value-added modeling is due to Hanushek (1971) and Sanders and Rivers (1996). For a more recent explanation of value-added

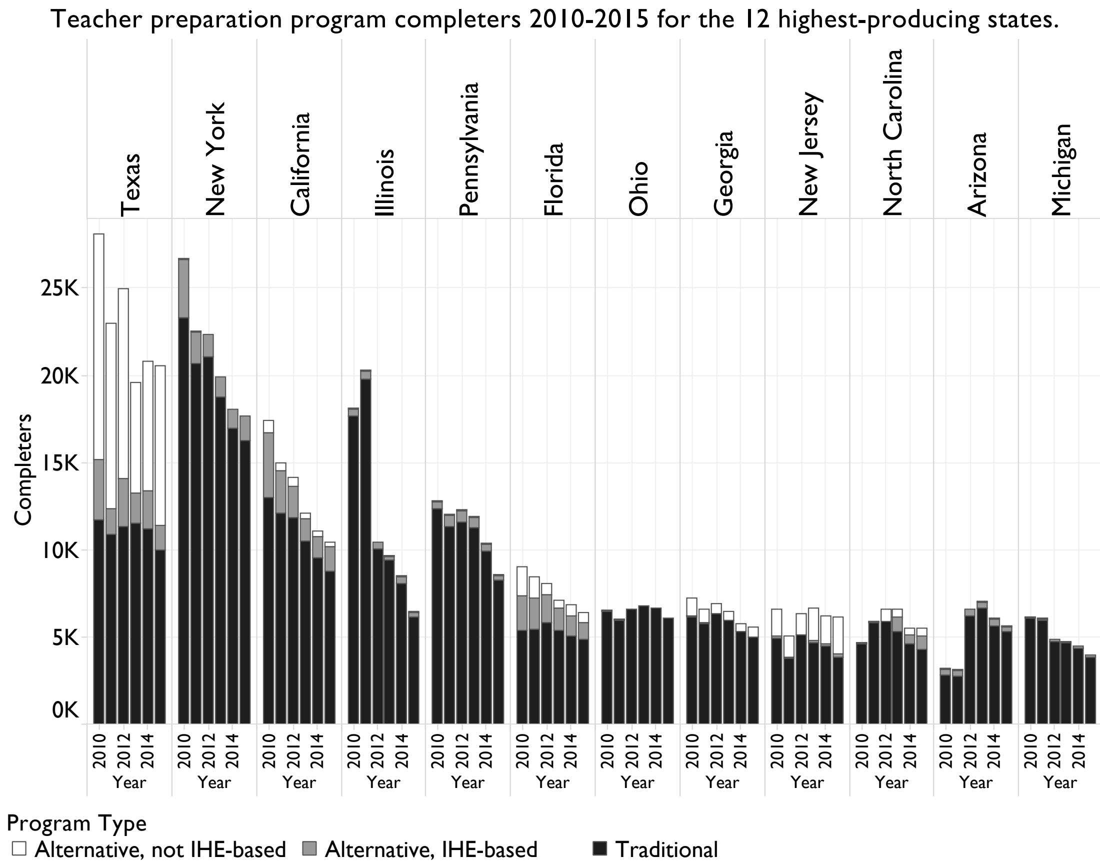


Figure 1: Teachers prepared by state over 6-year period (US Department of Education, 2016b, Completers 2010-2015). IHE refers to Institution of Higher Education.

procedures, see Rivkin et al. (2005), and for the economic benefits of improved teaching see Hanushek (2011). For a recent review see Koedel et al. (2015).

Important overviews of teacher certification pathways, including careful examination of evidence from studies of student learning, are due to Constantine et al. (2009), Grossman et al. (2008), Guarino et al. (2006) and Wayne and Youngs (2003).

There are several different lenses through which the literature on teacher preparation can be organized. Most studies focus on a particular geographical region, either a state or a district. Well-studied regions include Florida (Harris and Sass, 2011a; Sass, 2011), North Carolina (Henry et al., 2014b,a), Washington State (Goldhaber et al., 2013; Cowan and Goldhaber, 2016), Missouri (Koedel et al., 2012), New York City: (Kane et al., 2008; Boyd et al., 2007, 2009, 2012), California (Kane and Staiger, 2008), and Texas (Mellor et al., 2008; von Hippel et al., 2014).

Running through the literature is a collection of debates about technical points that challenge the validity of analysis. Most have to do with various sources of potential bias that reduce confidence in studies using administrative data as opposed to random-controlled trials. Aggregation bias may result when dissimilar units are grouped together (Hanushek, 2011). Grouping many units together can increase statistical power because of large sample size, but the differences may be due to unobserved factors rather than the distinctions defining the groups. Biased associations may be revealed by looking for apparently causal associations working backwards in time (Rothstein, 2009) (but see the reply from Goldhaber and Chaplin (2015); Koedel and Betts (2011) and Kinsler (2012)). Differences between preparation programs can be overstated because of not properly computing uncertainties given that students in different classrooms sometimes share the same teacher (Koedel et al., 2012).

Even when student score differences between regularly and alternatively certified teachers can be discerned, they are modest compared to the scale of differences set by standard deviation on the exams. The best-studied program in the country that recruits and supports alternatively certified teachers is Teach For America (TFA) (Decker et al., 2004; Turner et al., 2012; Clark et al., 2013). According to Clark et al. (2013), the difference in value-added effectiveness between TFA graduates and those of comparison programs is .06 standard deviations. These effects are measured in random controlled trials, and therefore have more internal validity than is possible from observational data. On the other hand, the schools in which the random controlled trials were conducted do not span a great range of school type so the external validity is limited. And there is also a literature challenging the effectiveness of TFA (Darling-Hammond et al., 2005; Laczko-Kerr and Berliner, 2002).

The literature on the specific relationship between preservice teacher preparation and subsequent student performance is not very large. Harris and Sass (2011b) review the literature up until 2011. Some studies, such as Kane et al. (2008) and Gordon et al. (2006) conclude that factors such as preparation routes and advanced degrees have almost no measurable effects on student outcomes, although there are large differences between individual teachers not captured by anything one knows of them before they begin to teach. Boyd et al. (2012), analyzing some of the same data from New York City as Kane et al. (2008), conclude that differences in teacher background can be detected; the difference in the analysis is mainly due to the attention given to program characteristics. The largest single effect in their base model is that a teacher have five years of experience, which corresponds to a value-added gain of 0.1 standard deviations in student test scores for middle school mathematics. The largest program differences, which are for Teach for America corps members, are around 0.05 standard deviations, while for College Recommended teachers the effect is around 0.02 standard deviations.

There are two studies in Texas prior to our current work particularly worth highlighting. Mellor et al. (2008) studied student learning outcomes in classrooms of beginning teacher graduates from University of Texas System campuses, with data from 2003 through 2007. Their primary goal was "to determine how student achievement in the classroom might be used as an indicator of the success of teacher preparation programs." At the time there was no statewide data system in place and they spent years obtaining data from over 400 districts. They carried out a variety of comparisons with multi-level models, but almost none of the effects they found was large enough to rule out having been caused by sampling uncertainty. They sum up by saying, "Our most significant finding was that limitations of most state data and assessment systems, including the one in Texas where our study was conducted, make this kind of research difficult." Six years later, the problem of evaluating learning gains due to Teacher Preparation Programs (TPPs) was revisited by von Hippel et al. (2014), now with the advantage of a statewide data set. They conclude, "In Texas we find that TPP estimates consist mostly of noise.... The potential benefits of TPP accountability may be too small to balance the risk that noisy TPP estimates will encourage needless, disruptive, and ineffective policy actions." These conclusions are similar to the findings of Koedel et al. (2012) in Missouri.

Concerning alternative certification the National Research Council concluded that "Because the information about teacher preparation and its effectiveness is so limited, high-stakes policy debates about the most effective ways to recruit, train, and retain a high-quality teacher workforce remain muddled."(NRC, 2010) Grossman et al. (2008, p. 185) similarly conclude that "[t]he available research does not paint a complete picture of either optimal recruiting and selection criteria nor optimal preparation opportunities." These studies are more than half a decade old, but conclusions have not changed much because "the field is moving in the direction of weighting value-added analyses in assessments of teacher preparation program quality (US Department of Education, 2016a, p. 487). For the purpose of active Federal policy, the point of view currently holding sway is that "effectiveness of graduates is not associated with any particular type of preparation program, [so] the only way to determine which programs are producing more effective teachers is to link information on the performance of teachers in the classroom back to their teacher preparation programs" (US Department of Education, 2016a, p. 566).

Alternative Certification and UTeach in Texas

Alternative Certification

Alternative certification of teachers was first permitted in Virginia in 1982, soon followed by California, Texas, and New Jersey (Suell and Piotrowski, 2007). Alternative certification is difficult to define precisely, and can encompass a wide range of programs, but in broad terms it describes "pathways designed to attract a wider range of candidates into teaching generally by reducing or eliminating pre-service education coursework and speeding paid entry into the classroom" (Grimmett and Young, 2012).

There are many potentially positive features of alternative certification.

- Reduce time and cost for candidates: For an individual who has completed a bachelor's degree, the time he or she will need before beginning paid teaching is much lower than the time that would be needed going through a university.
- Response to shortages: This raises the possibility of responding more quickly and efficiently to teacher shortages than universities can.
- Explore best practice: Alternative certification programs may be able to employ preparation coursework that is compressed in time and makes use of technology rather than face-to-face meetings. Successes achieved while exploring these possibilities could inform all preparation programs about how to operate more efficiently and effectively.
- Raise teacher quality: The introduction of alternative certification programs in a state may raise the caliber of person who enters teaching, by stripping away unnecessary, unappealing, or expensive coursework that previously constituted a barrier (Hess, 2001).

Yet in each of these respects alternative certification raises a concern.

- Someone who wishes to teach may have a financial interest in beginning paid teaching as soon as possible, but their students will not necessarily benefit if the new teachers are allowed to begin working full time before they are ready.
- The expansion of alternative certification programs could exacerbate teacher shortages by enabling reduction in the capacity of public institutions to prepare teachers.
- Cross-pollination of different types of programs could hardly be negative, but there is no evidence it has happened.
- The existence of alternative certification can reduce the motivation of undergraduates to obtain certification as part of their first degree, when it costs them nothing more than they would pay anyway to attend college. To obtain certification later requires hundreds or thousands of dollars more. The net result could be a decrease in the number or quality of individuals entering teaching.

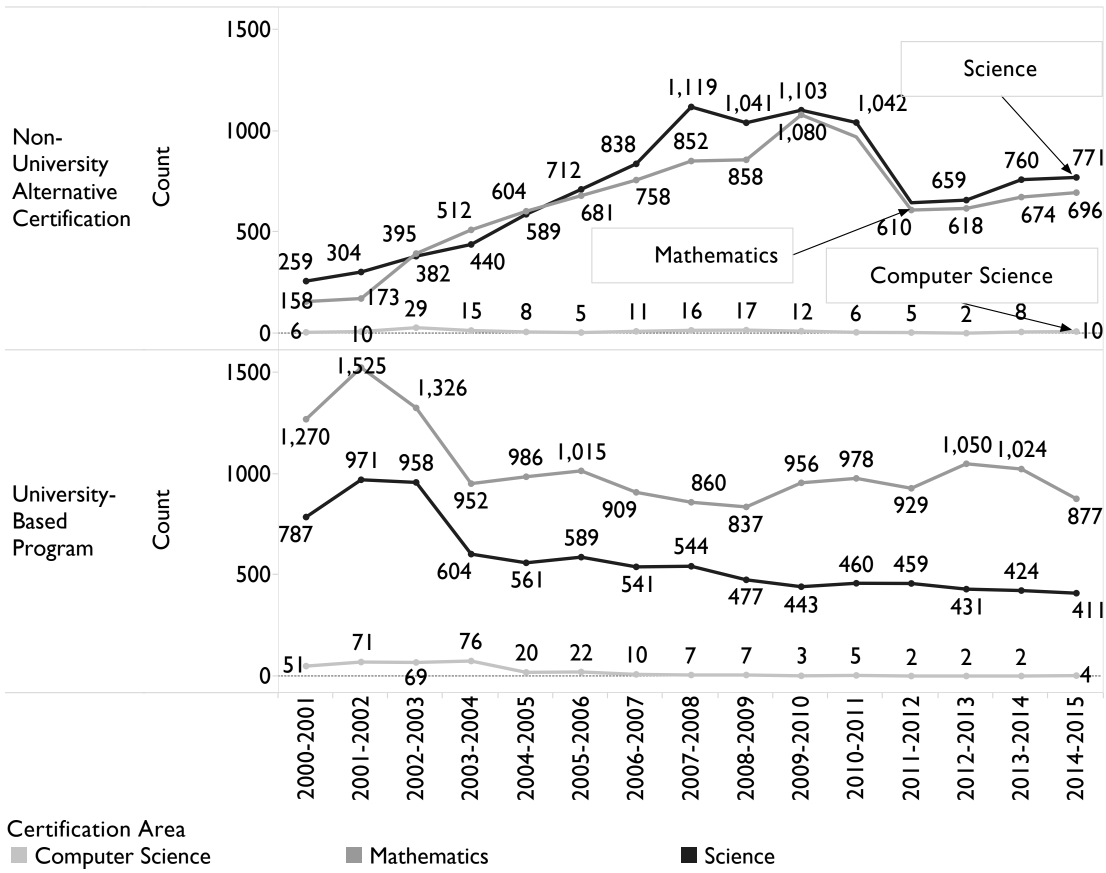


Figure 2: STEM teacher production in Texas from 2001 until 2015, comparing production from regular and alternative certification pathways.

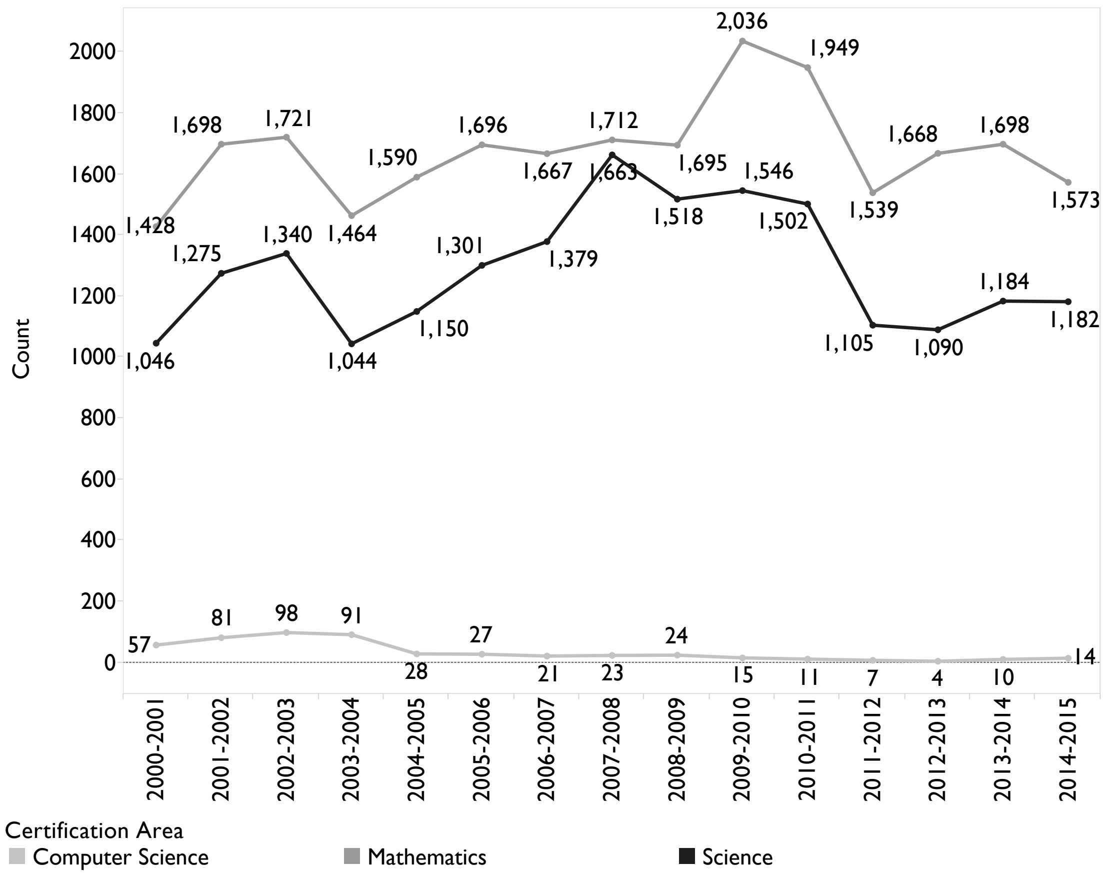


Figure 3: STEM teacher production in Texas from 2001 until 2015, summing all pathways.

Figure 2 shows the numbers of mathematics, science, and computer science teachers prepared in Texas since 2001 through university-based programs and non-university alternative certification. There are roughly three separate periods. Until 2003, universities clearly dominated the preparation of teachers. In 2004, university preparation fell by around 1/3 from a peak just two years before. At the same time, alternative certification began a period of rapid expansion, which continued until 2008, at which point it stabilized. In 2012, the number of teachers from alternative certification dropped by around 1/3 and has only slowly been recovering. The drop may be due to widely publicized cuts to school budgets in the spring of 2011, and to the fact that the economy was still recovering from the Great Recession. As shown in Figure 3, the total number of STEM teachers prepared in 2014-2015 was less than it had been in 2001-2002. The hope that alternative certification would suffice to eliminate teacher shortages has not been realized; on the other hand, this does not mean that Texas is worse off after the expansion of alternative certification than it would have been otherwise.

UTeach and Leading Texas Universities

UTeach began at the University of Texas at Austin in the fall of 1997 and since then has been the only UT Austin program certifying secondary STEM teachers. It is primarily a standard undergraduate certification program, although postbaccalaureate candidates can be recommended for certification by taking the same courses as the undergraduates. Features of UTeach include early field experience closely supervised by master teachers, active recruitment of content majors, coursework using research on how people learn, close cooperation between colleges of Natural Sciences and Education, and compact degree plans so students can get a degree and teaching certificate at the same time.

UTeach went through several phases. From 1998 to 2004 it grew to steady state at UT Austin. In 2007 UTeach started to be replicated across the United States, with three expansion sites in Texas. These sites began yielding graduates in 2010. As of 2016, UTeach has spread to 7 Texas institutions. These are UT Austin, UT Dallas, UT Arlington, UT Tyler, UT Rio Grande Valley (formerly UT Pan American and UT Brownsville), University of North Texas, and the University of Houston. The growth in number of graduates due to this expansion (Figure 4) has not yet reached steady state and should continue to increase.

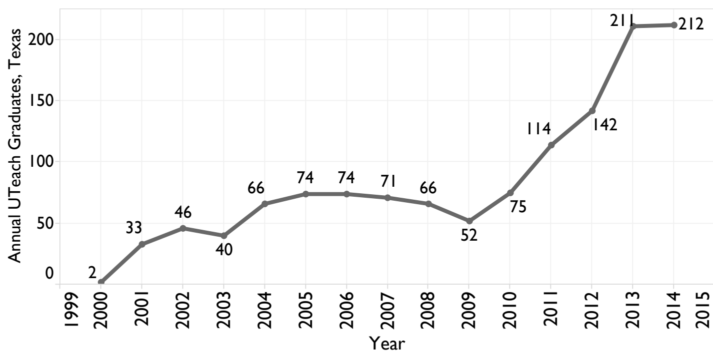


Figure 4: Texas STEM teachers prepared by UTeach

Despite this growth, UTeach is still a small enough program that we should expect to find little but noise if we heed the warnings of von Hippel et al. (2014) . However, the pedagogical practices within UTeach, including early field experience, employment of clinical faculty, and semester-long student teaching, are generally consistent with those of many other university-based teacher preparation programs. For example AggieTeach, the secondary preparation program at Texas A&M, which graduates more STEM teachers than any other university-based program in Texas, is quite similar to UTeach.

Thus, anticipating the difficulty of obtaining sufficient statistical power to detect effects when restricting attention to a small number of programs, we decided to perform some of the analysis by grouping together a collection of leading Texas universities that share similar practices. These are all UT System universities, all A&M System universities, Texas State University, University of Houston and University of North Texas. Specifically, we define these leading Texas universities to be University of Texas - Austin, University of Texas - Brownsville, University of Texas - Arlington, University of Texas - Dallas, University of Texas - Tyler, University of Texas - San Antonio, University of Texas - Pan American, University of Texas - El Paso Texas A&M University, Texas A&M University - Corpus Christi, Texas A&M University - Commerce, Texas A&M University - Kingsville, Texas A&M University - Texarkana, Texas State University-San Marcos, University of North Texas, and University of Houston

The logic behind choosing these universities is that they contain the state public research flagships, the largest former normal school, and all universities that eventually joined the UTeach network. The programs at these universities are not identical, nor are the populations of students, but the programs are by and large recognized as high quality, the students at the universities are among the state's best, and the programs have many things in common, including attention to content knowledge, pedagogical preparation, full student teaching, early field experience in many cases, and emphases on inquiry and equity. Thus they are similar enough to each other to be viewed as a group, and provide a large enough sample size that we are able to disaggregate to explore the space of variables described in Table 2.

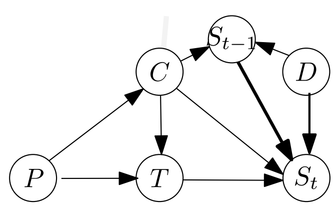
One of the problems particularly raised by the National Research Council (NRC, 2010) is that "there is more variation within categories such as 'traditional' and 'alternative' — and even within the category of master's degree programs — than there is between the categories." Based on our knowledge of particular programs, we believe this worry is less applicable to Texas than it may be to other jurisdictions. Both regular and alternative programs are governed by the same sets of rules: they must have 30 clock hours of field-work or observation, 300 clock hours of coursework, and either a student teaching semester or a year-long internship. However the traditions of the regular and alternative pathways are different. Alternative certification emerged from the philosophy that barriers to teaching should be removed, and the candidates, all of whom already have finished a first degree, usually have a few weeks of instruction and observation after which they enter the classroom working full time, completing their pedagogical coursework during an internship year. The university programs provide all the coursework, often but not necessarily as part of a degree, as well as the fieldwork prior to a student teaching semester, and only afterwards do the candidates begin to teach full time. To illustrate the point that alternative and standard programs differ in their practices, we note that for alternatively certified teachers of Algebra I in 2011-2012 with up to 10 years of experience, 97% entered teaching on a probationary certificate. This means they began teaching full time without having had student teaching. By contrast, for the Algebra I teachers from standard university programs, less than 1% entered on probationary certificates and the remaining 99% had a semester-long student teaching experience.

Causal Framework, Variables, and Sample

Causal Framework

Our view of how to draw causal inferences from data has been influenced by Pearl (2009) and Morgan and Winship (2015). One way to explain the approach is to say that we want to know how some intervention, such as a policy decision, might work out in the future. To estimate its effect, we look in the past for similar cases where such an intervention has operated and use those as a guide. The more similar we make the environment from the past to one we want to examine in the future, the greater the risk of drawing conclusions from small and noisy samples. Thus there has to be balance between grouping together such dissimilar things that they predict the future badly, and having such small samples that estimates end up swamped by uncertainty.

An idea due to Pearl (2009) is use of Directed Acyclic Graphs to describe causal relationships. These graphs display quantities that are correlated with each other, and indicate with arrows the direction that causality operates. We present such a diagram in Figure 5.



Symbol	*Meaning*
S_t	Student test score in year t.
S_{t-1}	Student test score in year $t-1$.
D	Demographic and other descriptive variables for student in year t.
T	Teacher assigned to student's class in year t.
C	Campus in year t.
P	Educator preparation pathway.

Figure 5: Causal diagram for effect of educator preparation pathway on student test scores.

For the most part, arguments about the direction of arrows in Figure 5 will have no obvious impact on the models we will write down in the end. However thinking about causality does influence the way model calculations are performed. For example, we do not control for teacher years of experience in our models, but we do look at two subsets of teachers: those with less than 4 years of experience, and those with less than 10 years of experience. The reason is that we are interested in the effect on student test scores of different teacher preparation pathways. Teachers coming from different pathways remain in teaching for different amounts of time (Figure 6). In particular, teachers from standard university programs stay at about a 10% higher rate than those from alternative certification programs. The population of novice teachers from universities is less heavily weighted with first-year teachers than novice teachers from alternative certification programs. Therefore, on the formal grounds that one should never control for the causal descendants of an intervention (Morgan and Winship (2015)) we do not control for years of teaching.

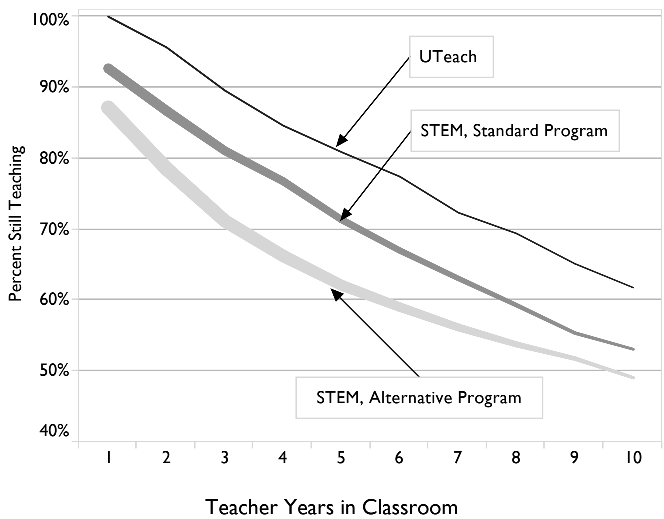


Figure 6: Retention of Texas STEM teachers in teaching by preparation pathway, averaged over cohorts entering from 2004 until 2013.

This argument makes it seem that we should never consider years of experience at all, but that is not right, either. States are required by US Department of Education (2016a) to hold TPPs accountable for their novice teachers, those with less than four years of experience. Programs will be rated effective or ineffective, and potentially stopped from preparing teachers, based on outcomes from this subgroup, so it has to be considered separately. Obviously it is also interesting to know how more experienced teachers perform. However, as shown in Figure 2, the Texas STEM teacher preparation landscape, with a balance of standard and alternative programs, has only had its current form since around 2002. Therefore, when looking beyond novice teachers to include those with more experience, we decided to restrict attention to those with up to 10 years of experience, so that the conditions under which they were prepared would reasonably resemble what we expect in the future.

Variables

We now provide a more detailed discussion of the symbols appearing in Figure 5. Each of the symbols expands into additional measured characteristics, which are not indicated in the diagram. Possible values of each variable appear in Table 2. They are Campus= C, Demographics= D, Test Scores=S and teacher=T. In each instance we have made choices about how to group variable values together.

Variable	Name	Category Values
C	School Campus	School Identifier *and* [Urban, Suburban, Small Town, Charter School]$\otimes$ School Poverty, Continuous $\in [0, 1]$ *or* Binned in four equal quartiles
D	Demographics	Flags for LEP, Free/Reduced Lunch, Gifted, Special Ed, and Race/Ethnicity
T	Teacher	Teacher Identifier *and* Years of Experience$\in$ [Novice$\equiv$<4 years experience, or $<$ 10 years experience]
S_t	Student raw test score in year t	Continuous $\in$ [0,1] or binned in $(0-.4], (.4-.5]\ldots(.9-1]$
P	Educator Preparation Pathway	[Alt, Post-Bacc, Standard, Out of State, Otheror [Post-Bacc/Standard, Alternative] *and* Texas UTeach, Leading Texas Universities, All]

Table 2: Variables used to stratify schools, classrooms, and teachers.

In the case of Campus C we characterize each campus by its type (Urban, Suburban, Small Town, Charter) and the school poverty concentration. The district types are described in Table 3. In most models we also include a varying intercept for each campus.

Texas Education Agency Category	***Group***
Major Urban	Urban
Other Central City	Urban
Major Suburban	Suburban
Independent Town	Small Town
Other Central City Suburban	Small Town
Non-Metropolitan Stable	Small Town
Rural	Small Town
Non-Metropolitan Fast Growing	Small Town
Charter School Districts	Charter

Table 3: Regrouping of Texas Education District types into the four categories used here.

For Demographics D we use student flags for race and ethnicity, special education, gifted, and limited English proficiency.

For Teacher T we use years of experience, sometimes restricted to novice teachers (defined as less than 4 years of experience), other times restricted to less than 10 years of experience.

For Test Scores S_{t-1} and S_t we used STAAR Algebra I in 2012 or STAAR Biology in 2012 for S_t and averages of the prior year 8th grade TAKS mathematics or 8th grade TAKS science for the same students to obtain S_{t-1}. 2011-2012 was the first year that student-teacher links became available in Texas on a statewide basis. The reason we use Algebra I and Biology is that they are the only high school exams in science and mathematics that remained after a substantial reduction of mandatory standardized testing from the Texas legislative session of 2013.

Finally, Figure 5 includes the effect of Educator Preparation Program P. As we have discussed, this appears as a fixed effect for program type (Alternative, Standard, or Postbacc with Standard Certificate) and then we restrict the Standard teachers to three program groupings: all university-based programs, programs at leading Texas universities, and Texas UTeach programs.

Sample Construction and Sample Size

We provide some information on the size of our sample in various categories. In Table 4 we show the numbers of teachers. We illustrate sample construction for the case of UTeach Algebra I teachers with less than 4 years of experience. We begin by finding all UTeach graduates with less than four years of experience teaching Algebra I. There are 48 of them. We find all the campuses in which they teach, and in those campuses we find all the alternatively certified teachers with less than 4 years of experience teaching Algebra I. That is the comparison group, labeled Alt in Table 4. There are 53 of them. The other comparison groups are constructed similarly.

Program Group	Subject	Years	Standard	Alt	Total
UTeach	Algebra I	<4	48	53	101
UTeach	Biology	<4	26	22	48
Leading	Algebra I	<4	268	192	460
Leading	Biology	<4	165	134	299
All	Algebra I	<4	694	802	1496
All	Biology	<4	503	887	1390
UTeach	Algebra I	<10	62	83	145
UTeach	Biology	<10	36	44	80
Leading	Algebra I	<10	432	460	892
Leading	Biology	<10	292	365	657
All	Algebra I	<10	1231	1348	2579
All	Biology	<10	959	1562	2521

Table 4: Numbers of teachers in study for all combinations of program group, pathway, years of experience, and discipline. The “Standard” column counts teachers from standard programs, while the “Alt” column counts the comparison group of alternatively certified teachers in the same schools.

Tables 5 and 6 provide descriptions of the numbers of students appearing in our sample in various categories. These are the numbers of students left after the exclusions described below under Weights and Exclusions. The categories Small Town, Suburban, and Urban, described in Table 3, have roughly equal numbers of students overall. The charter schools have many fewer, but as the Texas Education Agency breaks them out separately, and as they are of considerable policy interest, we keep them separate. Although around 300,000 students took the math and science exams, by the time we focus down on classrooms with a novice UTeach teacher, the numbers of students remaining may be only in the thousands, or in the case of charters schools in the dozens. Particularly for the charter schools, the numbers are not completely accurate because they were created by summing up smaller categories that sometimes had to be set to zero for masking.

Teacher assignments to Algebra I and Biology

Since we consider populations of teachers from different pathways teaching in different types of schools, it is interesting to know how they are associated. We provide some descriptive statistics addressing this matter in Figures 7 and 8. The numbers of teachers in the subcategories were sometimes less than five and subject to masking, so we report the numbers of students.

In mathematics, we note that the novice UTeach graduates are less likely than novice teachers from other program groups to have students in small towns, particularly high-poverty small towns, but are substantially more likely to have students in urban classrooms, particularly high-poverty urban classrooms. In science, this pattern is not repeated. Instead, the UTeach graduates are most likely to have students in wealthy suburban schools. It is not obvious why this is true since mathematics and science students in UTeach have an identical pedagogical preparation, with field experiences in low-income urban schools. Perhaps it is because the shortages described in Table 1 mean that the science teachers more than the mathematics teachers are recruited even as novices into schools with the greatest financial resources.

Methods

Overview and Restrictions

We present multilevel models where students are nested within a classroom, classrooms are nested within teacher, teachers are nested within campus, and we control for each student’s prescore, an array of demographic information about both student and campus, and estimate the effect of teacher pathway.

Discipline/Program Group/Pathway: <4 years experience

District	All: Alternative	All: Standard	Leading: Alternative	Leading: Standard	UTeach: Alternative	UTeach: Standard
Algebra I						
Charter	1766	770	29	317	10	106
Small Town	10664	10211	2115	3346	289	336
Suburban	15501	15672	3949	5537	1231	908
Urban	18561	13357	5375	6499	1442	1379
Total	46492	40010	11468	15699	2972	2729
Biology						
Charter	2355	716	191	342	.	31
Small Town	14675	8802	1646	2438	89	85
Suburban	21645	12766	4788	4205	1043	1101
Urban	21352	11014	3347	3809	336	1431
Total	60027	33298	9972	10794	1468	1431

Table 5: Numbers of students in study for combinations of discipline, program group, pathway, and district type. Teachers with < 4 years experience. The "Standard" column counts students with teachers from standard programs, while the "Alt" column counts the comparison group of students with alternatively certified teachers in the same schools.

We restrict the analysis in several ways. We used only data from the 2011-2012 academic year, with prescores from 2010-2011. The 2011-2012 year was the first year that student-teacher links are available in the Texas statewide dataset. In the present publication we have not considered additional years mainly as a practical matter. It took more than a year of work to clean the 2011-2012 dataset well enough that the results were useful. Policy changes in Texas concerning both high school testing and graduation requirements mean that results from different years are not necessarily easy to compare with each other even if the cleaning process is complete. Thus, we leave analysis of later years for later work.

"Months of Schooling" Units, (MOS)

It is common to define an effect size by dividing exam scores by the standard deviation. For ninth graders who took Algebra I in 2012-2013, the standard deviation was .17, while for these same students the standard deviation of their mathematics scores the year before in 8th grade was .15. Therefore we take 0.16 as the standard deviation. We define "Months of Schooling" units where 9 Months of Schooling (MOS) corresponds to one quarter of a standard deviation (Gates Foundation, 2012). In these units, gaining 0.04 in raw score on an exam is reported as 9 Months of Schooling. Note that in these units, for both Algebra and Biology, the difference in expectations between the lowest and highest poverty concentration classrooms is on the order of nine months of schooling. Also note that one month of schooling corresponds to .028 or 2.8% in standard deviation units. Thus, the effects found in previous studies in New York City or for Teach for America are on the order of 1 to 2 in these Months of Schooling units. To convert from percentage of a standard deviation to Months of Schooling, multiply by 0.36. To convert from raw score to Months of Schooling, multiply by 225. Although one added Month of Schooling sounds like a large difference, it is a small effect. There are around 50 problems on the Texas exams, and it corresponds to one chance in five of getting one more problem right.

Discipline/Program Group/Pathway: <10 years experience

Algebra I

District	All: Alternative	All: Standard	Leading: Alternative	Leading: Standard	UTeach: Alternative	UTeach: Standard
Charter	2527	969	115	330	.	106
Small Town	17946	18722	5160	5798	377	364
Suburban	24570	25777	10241	8915	1773	1359
Urban	28949	23990	10975	9522	2144	1601
Total	73992	69458	26491	24565	4294	3430

Biology

District	All: Alternative	All: Standard	Leading: Alternative	Leading: Standard	UTeach: Alternative	UTeach: Standard
Charter	3598	949	207	442	.	31
Small Town	24305	18159	4864	4948	224	285
Suburban	39264	22105	11927	7007	2150	1645
Urban	36125	20909	9293	7045	608	406
Total	103292	62122	26291	19442	2982	2367

Table 6: Numbers of students in study for various combinations of discipline, program group, pathway, and district type. Teachers with <10 years experience. The "Standard" column counts students with teachers from standard programs, while the "Alt" column counts the comparison group of students with alternatively certified teachers in the same schools.

Groups

We found during preliminary investigations that there is a strongly nonlinear interaction between poverty concentration and district type, and pretest scores. For this reason we defined a variable G that creates comparison groups with

$$\begin{aligned} G &= \text{District} \in \{\text{Urban, Suburban, Small Town,Charter}]\} \\ &\otimes \text{Prescore}{=}S_{t-1} \in 0-.4](.4-.5],(.5-.6]\ldots(.9-1] \\ &\otimes \text{School Poverty Concentration} \in \{[0-25\%],(25\%-50\%],(50\%-75\%],(75\%-100]\} \end{aligned} \tag{1}$$

The use of this variable in models enables us to control for the interactions of district type, prescore, and poverty concentration in a non-parametric fashion. We also explored a wide range of parametric representations of these and other factors.

Weights and Exclusions

Any given student test score result could end up in our data set from one to six times. The test scores appeared multiple times when the student took classes with separate identification numbers in separate semesters, and when more than one teacher was associated in the data with the class section. We weighted every student record inversely with the number of times the student appeared, so if a student was taught by several teachers during the year, each of them shared equally, and that student did not contribute more to the final results than a student who appeared only once.

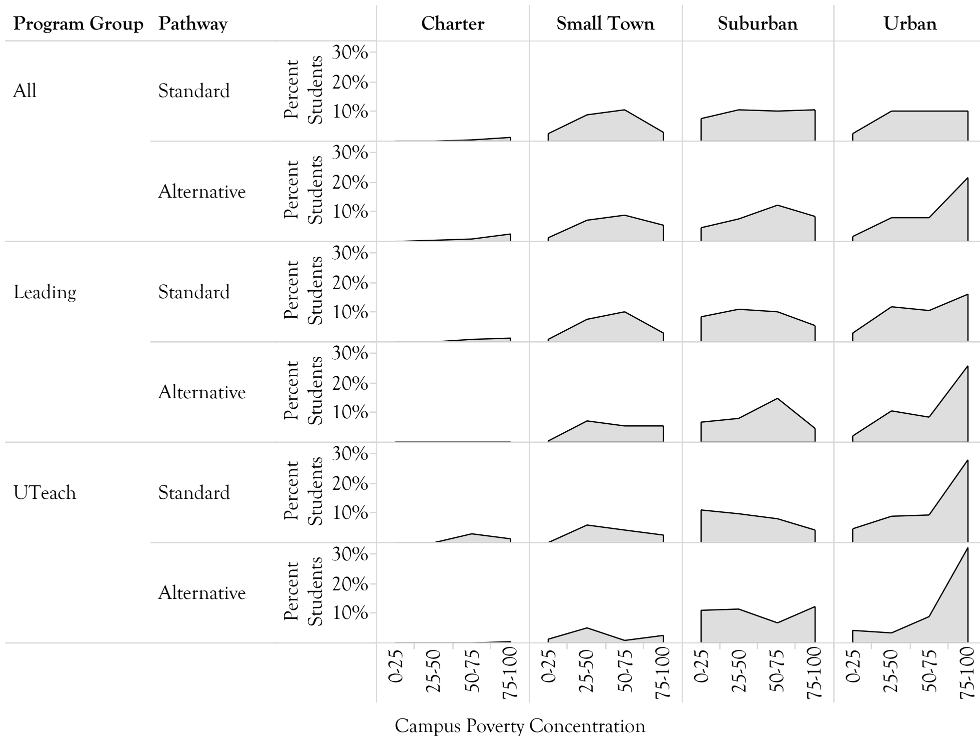


Figure 7: How the sample for Algebra I students taught by teachers with less than 4 years experience is distributed in different types of school environments. In every row the percentage of students adds to 100%.

We excluded many records. We kept only cases where the student had a valid ninth-grade score in 2012 and a valid eighth-grade score in 2011. There were thousands of students who took Algebra I in eighth grade in 2012; this is quite a different population than the ninth-graders and we do not report the results here. There were several accommodations available to students both in 2011 and 2012, including provisions for English-language learners, vision-impaired students, and a modified exam for students with learning disabilities. In most of our analyses, we exclude all students who received any of these accommodations in either 2011 or 2012. However, in our report on student-level models, we do at one point include results from students who received an accommodation in 2011 and also received one in 2012; most of them were taking the alternate exam, so they cannot be compared easily with the other students.

We also limited the groups of teachers we examined. The state has many different categories, including those who were prepared out of state. We limited ourselves to teachers from standard programs entering teaching with a standard certificate, teachers from university postbaccalaureate programs receiving a standard certificate (these two sets of university-prepared teachers were combined in the final analysis), and we compared them with teachers from alternative certification programs who began teaching with a provisional certificate. We excluded postbaccalaureate candidates from universities who entered teaching with provisional certificates, since by not entering with a standard certificate we deduce they did not have a student teaching semester and thus their preparation pathway differs substantially from the standard one.

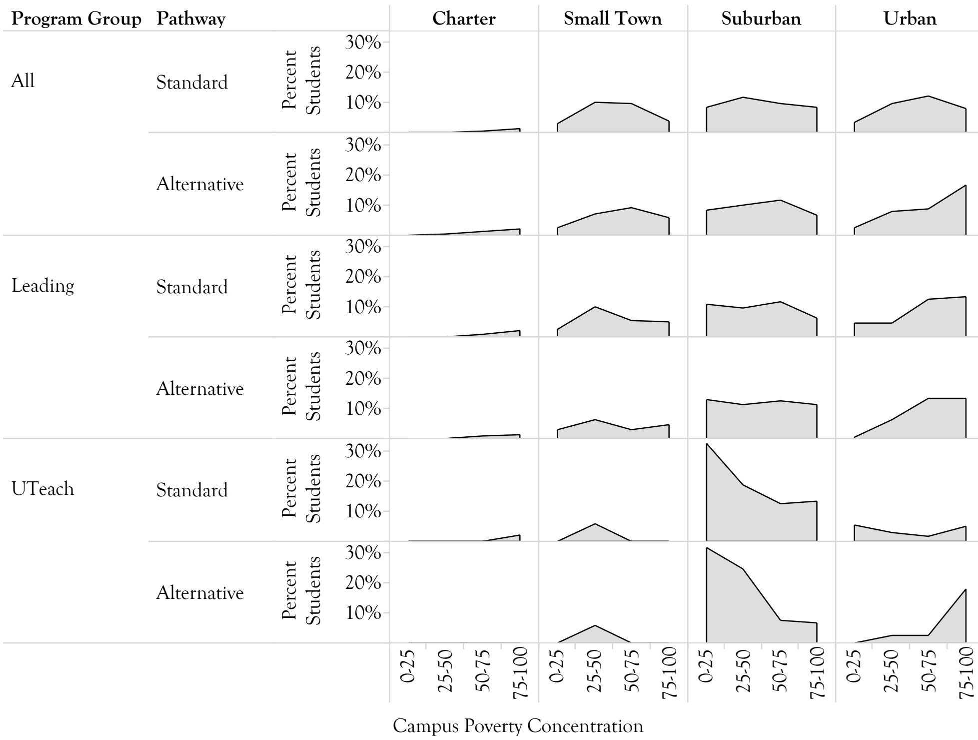


Figure 8: How sample for Biology students taught by teachers with less than 4 years experience is distributed in different types of school environments. In every row the percentage of students adds to 100%.

Multilevel Models

We explored a number of different multilevel models using lmer in R (Bolker et al., 2016). The one we view in the end as our baseline model has the form

$$S_{i,t} \sim N(\sum_{\beta=1}^{3} \lambda_\beta S_{i,t-1}^{\beta} + T_{j[i]} + C_{k[i]} + \mathrm{Class}_{n[j[i]]} + \mathrm{Cert}_{m[j[i]]} + \sum_X X_{g[i]}; \sigma_S^2) \tag{2}$$

at the top level, where $S_{i,t}$ is the score of student i in year t and $S_{i,t-1}$ is the student's score on the exam in the same subject the previous year in a cubic polynomial. The cubic polynomial is needed because the relationship between current and prior year score is very nonlinear; studies such as Cowan and Goldhaber (2016) and Boyd et al. (2009) also model prior score up to the cubic level. The coefficients for demographic factors X range over Gifted, racial and ethnic groups, Limited English Proficiency (LEP), Free/Reduced Lunch Eligibility, and Special Education; here $g[i]$ is the value of group membership for student i. Certification pathway Cert_m enters as a fixed effect. By modeling the teacher in this way, each teacher should contribute equally to the estimate of the effect of their pathway to teaching. The second level of the model has random intercepts for teacher T, campus C, and class section Class,

$$T_j \sim N(\mu_T; \sigma_T^2) \quad C_k \sim N(\mu_C; \sigma_C^2) \quad \mathrm{Class}_n \sim N(\mu_L; \sigma_L^2).$$

A second variant of the model is

$$S_{i,t} \sim N(\sum_{\beta=1}^{3} \lambda_\beta S_{i,t-1}^{\beta} + T_{j[i]} + C_{k[i]} + \mathrm{Class}_{n[j[i]]} + \mathrm{Cert}_{m[j[i]]} + \sum_X X_{g[i]}; \sigma_S^2) \tag{3}$$

$$T_j \sim N(\mu_T; \sigma_T^2) \quad \mathrm{Class}_n \sim N(\mu_L; \sigma_L^2).$$

This is the same, except that campus is treated as a fixed effect at the top level, rather than being modeled as a random effect at the second level. While this model more accurately captures the effect of each campus, it is less appropriate for finding the contribution of teacher pathway because, for example, in cases where a campus has teachers from only a single pathway, the campus fixed effect subtracts off any effect these teachers have rather than comparing them with teachers in similar campuses as the campus random effect model does.

A third variant of the model is

$$S_{i,t} \sim N(G_{l[i]} + T_{j[i]} + C_{k[i]} + \mathrm{Class}_{n[j[i]]} + \mathrm{Cert}_{m[j[i]]} + \sum_X X_{g[i]}; \sigma_S^2) \tag{4}$$

In this case, rather than a polynomial function of prescore, we use the groups described by Eq. 1, composed from all combinations of campus poverty quartiles and prescore deciles, modeling teacher, campus, and class section with random intercepts at the second level.

$$T_j \sim N(\mu_T; \sigma_T^2) \quad C_k \sim N(\mu_C; \sigma_C^2) \quad \mathrm{Class}_n \sim N(\mu_L; \sigma_L^2).$$

We had one more variant, which we applied to subpopulations.

$$S_{i,t} \sim N(\sum_{\beta=1}^{3} \lambda_\beta S_{i,t-1}^{\beta} + T_{j[i]} + C_{k[i]} + \mathrm{Class}_{n[j[i]]} + \mathrm{Cert}_{m[j[i]]}; \sigma_S^2) \tag{5}$$

This equation was limited, for example to the subpopulation of economically disadvantaged students, or to gifted students.

Results

Random Effects		Fixed Effects	
Group	Standard Deviation (MOS)	Group	Effect (MOS)
Campus	11	Econ Dis	-2.2 ± 0.1
Teacher	7.4	Black	-3.0 ± 0.2
Class	2.5	Hispanic	-3.7 ± 0.2
		Asian	5.6 ± 0.4
Fixed Effects for Pretest Score (MOS)		Gifted	9.0 ± 0.3
λ_1	48.4 ± 12.6	Special Ed	-2.8 ± 0.5
λ_2	-9.59 ± 19.5	LEP	-2.5 ± 0.2
λ_3	79.7 ± 9.8	Standard Cert	1.15 ± 0.4

Table 7: Coefficients for random and fixed effects for Algebra I teachers with less than 10 years of experience, model from Eq. 2.

We present results from our multi-level models. We select one to examine in detail and explain the relative size of various effects. Consider the model in Eq. 2 for all teachers with less than 10 years of experience. The random and fixed effects are given in Table 7 for Algebra I and Table 8 for Biology. The coefficients for the models of Algebra I and Biology are quite similar to each other.

Among the random effects, the largest is the difference between campuses, with a standard deviation of 11 Months of Schooling for Algebra I and 10 in Biology. The standard deviation of the difference between teachers is around 7 Months of Schooling in both subjects. Thus we find larger differences between campuses than within them. The standard deviation of classes taught by the same teacher is around 3 Months of

Random Effects		Fixed Effects	
Group	Standard Deviation (MOS)	Group	Effect (MOS)
Campus	9.8	Econ Dis	-2.5 ± 0.1
Teacher	7.7	Black	-1.2 ± 0.2
Class	3.2	Hispanic	-4.5 ± 0.2
		Asian	6.1 ± 0.3
Fixed Effects for Pretest Score (MOS)		Gifted	11.5 ± 0.2
λ_1	193.2 ± 9.1	Special Ed	-3.7 ± 0.5
λ_2	-349.5 ± 14.1	LEP	-4.2 ± 0.2
λ_3	281.0 ± 7.0	Standard Cert	0.7 ± 0.4

Table 8: Coefficients for random and fixed effects for Biology teachers with less than 10 years of experience, model 2.

Schooling in both subjects. Each of these effects is larger than the effect due to teacher pathway, which is around 1 Month of Schooling.

Figure 9 provides a graphical overview of our results for each combination of teacher pathway, subject, years of experience, and a variety of student subgroups, using the model in Eq. 2. Almost all of the estimates, overall or by subgroup, indicate that students of teachers with standard certification score higher by around one Month of Schooling than students in the same schools and subjects whose teachers were alternatively certified. The largest difference appears to be for students flagged as Gifted, but the results from Economically Disadvantaged students and those of Limited English Proficiency are also noteworthy.

In Table 9 we provide overall estimates for each combination of pathway, subject, and years of experience for the three models in Eqs. 2-4. While every single one of these estimates is positive, the only ones to meet simple tests of statistical significance are for Leading Universities and All Universities in Algebra I. From the perspective of providing guidance for the future, estimating the effect of All Universities this way makes more sense than estimating the effects of Leading Universities or UTeach. The reason is that the comparisons are limited to the schools where the teachers in our sample are teaching. One sees in Figures 7 and 8 that the Leading University and UTeach teachers concentrate in school environments that are not representative of Texas as a whole. Following the prescription of Pearl (2009), Eq. 3.13, one should construct an overall estimate by summing over sub-populations weighted according to their expected presence in future interventions. This is not possible, because the future interventions are not known. However we make two comments. First, we observe that a reweighting procedure is probably not called for in the case of the All Universities sample for teachers with up to 10 years of experience, since this already includes the bulk of the teachers in almost every high school in the state. Second, we computed the effect of teacher pathway on students belonging to a variety of subpopulations. Thus, for example, if there is to be an intervention that directs new teachers to work in schools with a majority of low-income students, as Noyce Scholarships and TEACH grants do, one can try to assess the impact on populations policy is aiming to affect.

In Tables 10 and 11 we report model estimates for subpopulations of students using Eq. 5. For Algebra I, 37 of the 42 estimates are positive, favoring teachers with standard certification, 10 of the estimates are statistically significant, and all favor teachers with standard certification. For Biology, 34 of the 42 estimates are positive, 7 of the estimates are statistically significant, and all of these favor teachers with standard certification. For the Students column labeled "Mod" we look at results from Special Education students who took an alternative, modified, exam in both 2011 and 2012. In all other cases, we examine results from the regular exam restricted to a subpopulation.

As emphasized by von Hippel et al. (2014), the estimates are quite noisy. We provide a simple expression to estimate the uncertainty. Focusing on the results of Eq. 2, the uncertainty is fit quite accurately by

$$\text{Standard Uncertainty (MOS)} \approx \frac{17}{\sqrt{N_{\text{Teachers}}}}, \tag{6}$$

where N_{Teachers} is the sum of the number of standard and alternatively certified teachers being compared; fitting this expression to the uncertainty produced by the multilevel model gives R^2 of 0.98. What this implies is that to detect a difference of 1 month of schooling between two groups of teachers, there needs to be a sample of around 1000 teachers. The UTeach sample is roughly three times smaller than the Leading University sample, which in turn is four times smaller than the sample from All Universities. The scale for the uncertainty in Figure 9 grows accordingly as one moves from the largest program grouping to the smallest. It is a little surprising that the models produce any statistically significant estimates for UTeach at all; these are for Gifted students in mathematics, and for Gifted, Economically Disadvantaged, and Hispanic students in science.

We recall that in previous studies such as (Boyd et al., 2012) a change of 0.05 standard deviations was the largest typically ascribed to a group of teachers; in current units that corresponds to 1.8 Months of Schooling. Thus, although the effects we find in Tables 9 – 11 are small compared to the standard deviation of teachers or schools, they are consistent with teacher pathway effects found previously.

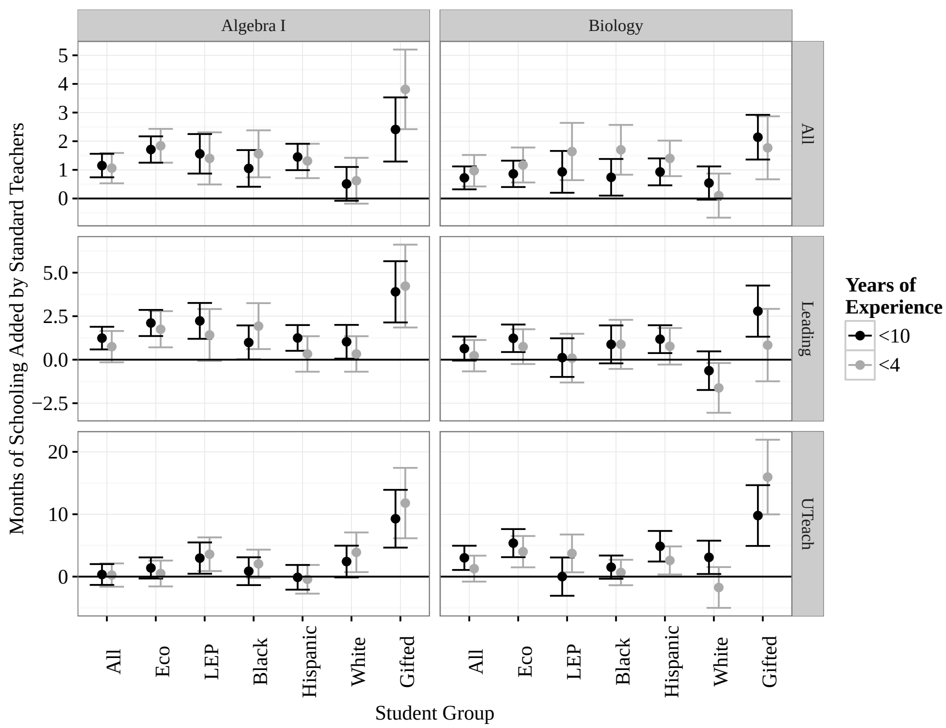


Figure 9: Estimates of added Months of Schooling for ninth-grade Algebra I and Biology students of novice and experienced teachers from all standard Texas university programs, leading Texas universities, and UTeach universities, using the model in Eq. 2. Columns provide separate estimates for a variety of student subgroups. Students receiving accommodations are excluded. Note that vertical scales are not the same for all rows. Error bars indicate one standard uncertainty.

Pathway, Subject	UTeach Algebra I					
Effect (MOS)	0.25±1.87	0.90±1.84	0.32±1.97	0.34±1.67	1.07±1.71	1.23±1.74
Years	4			10		
Pretest Control	C	N	C	C	N	C
Campus Intercept	R	R	F	R	R	F
Pathway, Subject	UTeach Biology					
Effect (MOS)	1.28±2.08	2.48± 2.29	1.07±2.13	3.01±1.94	4.31±1.99 *	3.61±2.02
Years	4			10		
Pretest Control	C	N	C	C	N	C
Campus Intercept	R	R	F	R	R	F
Pathway, Subject	Leading Texas Universities, Algebra I					
Effect (MOS)	0.73±0.83	0.88±0.94	0.86±0.93	1.24±0.65	1.54±0.67 *	1.3±0.68
Years	4			10		
Pretest Control	C	N	C	C	N	C
Campus Intercept	R	R	F	R	R	F
Pathway, Subject	Leading Texas Universities, Biology					
Effect (MOS)	0.23±0.90	0.56±0.94	0.57±0.94	0.64±0.69	0.65±0.72	1.18±0.72
Years	4			10		
Pretest Control	C	N	C	C	N	C
Campus Intercept	R	R	F	R	R	F
Pathway, Subject	All Texas Universities, Algebra I					
Effect (MOS)	1.06±0.53 *	0.95±0.54	0.74±0.61	1.15±0.41 **	1.00±0.42 *	0.78±0.45
Years	4			10		
Pretest Control	C	N	C	C	N	C
Campus Intercept	R	R	F	R	R	F
Pathway, Subject	All Texas Universities, Biology					
Effect (MOS)	0.97±0.55	0.71±0.56	0.35±0.66	0.72±0.4	0.45±0.42	0.68±0.45
Years	4			10		
Pretest Control	C	N	C	C	N	C
Campus Intercept	R	R	F	R	R	F

Table 9: Overall estimates for effects of teachers with standard certificates. We report estimates for all combinations of teacher pathway, subject, and cutoff on teacher years of experience. In each case we report on three models. The first is Eq. 2, which controls for pretest score with a cubic polynomial C, and models each campus as a random factor R. The second uses Eq. 4 and substitutes the non-parametric control for prescore and campus type N described in Eq. 1. The particular values reported here come from using prescore deciles, but using prescore quartiles only affects the second decimal place. The third model, Eq. 3, returns to polynomial control for prescore but includes each campus as a fixed effect. We estimate statistical significance by using * for $|t| > 1.96$, ** for $|t| > 2.58$, where t is the ratio of effect to uncertainty.

Teachers	Years	Students	Effect (MOS)	Significant
All	<4	Mod	-0.44 ± 1.23	
All	<10	Mod	1.39 ± 0.89	
Leading	<4	Mod	-2.46 ± 2.35	
Leading	<10	Mod	-0.30 ± 1.62	
UTeach	<4	Mod	3.80 ± 6.50	
UTeach	<10	Mod	6.75 ± 5.75	
All	<4	LEP	1.40 ± 0.91	
All	<10	LEP	1.56 ± 0.69	*
Leading	<4	LEP	1.42 ± 1.49	
Leading	<10	LEP	2.23 ± 1.03	*
UTeach	<4	LEP	3.60 ± 2.60	
UTeach	<10	LEP	2.97 ± 2.51	
All	<4	Gifted	3.81 ± 1.39	**
All	<10	Gifted	2.41 ± 1.12	*
Leading	<4	Gifted	4.23 ± 2.38	
Leading	<10	Gifted	3.90 ± 1.76	*
UTeach	<4	Gifted	11.8 ± 5.60	*
UTeach	<10	Gifted	9.28 ± 4.63	
All	<4	Eco	1.84 ± 0.59	**
All	<10	Eco	1.71 ± 0.46	**
Leading	<4	Eco	1.75 ± 1.04	
Leading	<10	Eco	2.11 ± 0.75	**
UTeach	<4	Eco	0.50 ± 2.10	
UTeach	<10	Eco	1.39 ± 1.69	

Teachers	Years	Students	Effect (MOS)	Significant
All	<4	White	0.62 ± 0.80	
All	<10	White	0.51 ± 0.59	
Leading	<4	White	0.33 ± 1.02	
Leading	<10	White	1.03 ± 0.97	
UTeach	<4	White	3.90 ± 3.50	
UTeach	<10	White	2.42 ± 2.54	
All	<4	Black	1.56 ± 0.82	
All	<10	Black	1.05 ± 0.64	
Leading	<4	Black	1.93 ± 1.32	
Leading	<10	Black	0.99 ± 0.98	
UTeach	<4	Black	2.00 ± 2.20	
UTeach	<10	Black	0.87 ± 2.23	
All	<4	Hispanic	1.31 ± 0.60	*
All	<10	Hispanic	1.45 ± 0.46	**
Leading	<4	Hispanic	0.33 ± 1.02	
Leading	<10	Hispanic	1.25 ± 0.74	
UTeach	<4	Hispanic	-0.42 ± 2.30	
UTeach	<10	Hispanic	-0.11 ± 1.98	

Table 10: Algebra I: Multilevel model estimates for effect of teacher pathway on subgroups of students. The effects come from Eq. 5 applied to a restricted population of students.

Teachers	Years	Students	Effect (MOS)	Significant
All	<4	Mod	-0.03 ± 1.01	
All	<10	Mod	0.17 ± 0.74	
Leading	<4	Mod	-0.38 ± 1.82	
Leading	<10	Mod	-0.63 ± 1.33	
UTeach	<4	Mod	-6.30 ± 4.15	
UTeach	<10	Mod	-6.40 ± 3.30	
All	<4	LEP	1.64 ± 1.00	
All	<10	LEP	0.93 ± 0.73	
Leading	<4	LEP	0.09 ± 1.40	
Leading	<10	LEP	0.12 ± 1.11	
UTeach	<4	LEP	3.72 ± 3.03	
UTeach	<10	LEP	0.00 ± 3.07	
All	<4	Gifted	1.77 ± 1.10	
All	<10	Gifted	2.14 ± 0.78	**
Leading	<4	Gifted	0.84 ± 2.08	
Leading	<10	Gifted	2.79 ± 1.47	
UTeach	<4	Gifted	15.96 ± 5.98	**
UTeach	<10	Gifted	9.79 ± 4.87	*
All	<4	Eco	1.17 ± 0.61	
All	<10	Eco	0.86 ± 0.46	
Leading	<4	Eco	0.75 ± 1.00	
Leading	<10	Eco	1.23 ± 0.79	
UTeach	<4	Eco	4.00 ± 2.51	
UTeach	<10	Eco	5.37 ± 2.25	*

Teachers	Years	Students	Effect (MOS)	Significant
All	<4	White	0.10 ± 0.77	
All	<10	White	0.54 ± 0.58	
Leading	<4	White	-1.62 ± 1.43	
Leading	<10	White	-0.63 ± 1.11	
UTeach	<4	White	-1.74 ± 3.27	
UTeach	<10	White	3.09 ± 2.67	
All	<4	Black	1.70 ± 0.87	
All	<10	Black	0.74 ± 0.64	
Leading	<4	Black	0.88 ± 1.41	
Leading	<10	Black	0.88 ± 1.09	
UTeach	<4	Black	0.66 ± 2.04	
UTeach	<10	Black	1.52 ± 1.86	
All	<4	Hispanic	1.40 ± 0.62	*
All	<10	Hispanic	0.93 ± 0.47	*
Leading	<4	Hispanic	0.77 ± 1.05	
Leading	<10	Hispanic	1.18 ± 0.80	
UTeach	<4	Hispanic	2.59 ± 2.25	
UTeach	<10	Hispanic	4.87 ± 2.45	*

Table 11: Biology: Multilevel model estimates for effect of teacher pathway on subgroups of students. The effects come from Eq. 5 applied to a restricted population of students.

Conclusions

Students of teachers from standard pathways in Texas have better high school math and science learning outcomes than students of teachers from alternative pathways. In Algebra I students of standard teachers gain 1 Month of Schooling overall. In Biology the effect appears to be around 0.7 Months of Schooling, although statistical uncertainty is too great for us to make conclusive statements about the Biology students overall.

We can speak more confidently about the student outcomes when we focus on specific populations. For Algebra I, teachers with standard certificates obtain significantly better results for students with Limited English Proficiency, Gifted students, Economically Disadvantaged students, and Hispanic students. In Biology, teachers with standard certificates obtain significantly better results with Gifted students, Economically Disadvantaged students, and Hispanic students. The size of the effect varies according to the preparation program group and years of experience, but the estimates are of differences as high as 5 to 9 Months of Schooling. The only subgroup for which alternatively certified teachers may have an edge is for Special Education students taking modified exams. Most of the estimates favor alternatively certified teachers for this population, although the effects are a bit below conventional thresholds for statistical significance.

Some prior studies have concluded that characteristics of teacher education are too small to detect or too small to matter in student achievement (Gordon et al., 2006; Rivkin et al., 2005; Staiger and Rockoff, 2010; Aaronson et al., 2007; Harris and Sass, 2011b). We find that teachers with standard certification get better results, particularly in mathematics, and this is true both for novice and experienced teachers. However it takes a large number of teachers and students to obtain results. One reason, seen in Tables 7 and 8 is that for a single teacher teaching multiple sections of the same class at the same time, the typical variation from one section of the class to another is around 3 Months of Schooling. To estimate the effect of any particular type of teacher preparation pathway with uncertainty less than 1 Month of Schooling, one must average the results of around 1000 teachers. We accomplished this by aggregating together preparation programs in groups with similar practices rather than trying to find effects at the level of a single program. Such grouping also is a feature of the finding of positive associations between National Board certification and student achievement in Cowan and Goldhaber (2016), and of pathways in New York City by Boyd et al. (2009).

It has frequently been stated that that variance between classrooms in schools is larger than variance between schools (Nye et al., 2004). Staiger and Rockoff (2010) conclude that "School leaders have very little ability to select effective teachers during the initial hiring process" and present as evidence "the fact that most of the variation in teacher effects occurs among teachers hired into the same school." We did not find this. In Table 7, variation between schools is the largest random effect, followed by variation between teachers in schools, then followed by different classrooms of the same teacher. Thus the question of whether some school leaders are able to select effective teachers should remain open.

We recover a result, often found before, that there is more variation of student outcomes within teacher preparation pathways than between preparation pathways. This finding has been used in support of policies that reduce barriers for new people to enter teaching, but make it difficult for them to continue unless they can demonstrate favorable student outcomes (Gordon et al., 2006). While such policies might make sense in cases where there are more people wishing to become teachers than there are positions available, they are harder to justify for shortage areas such as secondary STEM. It is hard to imagine that either young people or career changers will be attracted to secondary teaching by the prospect of high-stakes evaluations coming from multi-level models operating on their students' test scores; newspaper accounts such as those of Bonner (2016) conclude that these evaluations have been exacerbating teacher shortages. Teacher shortages may not directly impact high-stakes subjects such as Algebra I and Biology; schools have to staff them or face severe penalties. Shortages show up more naturally for subjects such as Computer Science where there are no high-stakes assessments, and where only a small fraction of high schools even offers a course (Guzdial, 2012). It is tempting to consider policies that make it difficult for teachers with low value-added scores to continue teaching, in hopes of capturing some of the 7 Months of Schooling advantage for the best teachers in Table 7. However, reducing the stability of teaching careers will impact the individuals who decide to enter teaching, and there is no assurance that secondary students will benefit in the end.

Our results do not justify an abrupt policy change impacting alternative certification programs in Texas. We found differences between standard and alternative programs, but they are not extremely large. In mathematics there are many subgroups of students for whom the advantages of having a teacher from a standard university program are significant; in science the cases where we found significance are fewer. Once again, one must keep in mind because of the shortage of STEM teachers that it is difficult to justify reducing teachers from any pathway. Slightly increasing the scores of low-income students on Biology exams but reducing the number able to take Physics or Chemistry at all would almost certainly be a very poor trade. On the other hand, our results do not provide strong incentive for other states to follow Texas's lead in establishing a large for-profit alternative certification sector. As shown in Figures 2 and 3, the growth of alternative certification in Texas since the early 2000s has not led in the end to an increase in the production of STEM teachers.

Overall, we found that teachers prepared by standard programs stay in teaching longer than those from alternative certification and their students learn more. The learning gains due to teachers from standard programs are most pronounced for groups such as Hispanic, Economically Disadvantaged, and Gifted students. Around 700,000 undergraduates obtain STEM degrees from US universities each year. This is an enormous pool; persuading just 1% more to obtain a teaching certificate along with their degree each year would add 7000 new STEM teachers. Thus, we encourage support for the preparation of STEM teachers through standard university pathways as an efficient, scalable, high-quality way to address the critical need for improved STEM education and to address the shortage of STEM teachers.

Acknowledgements

This work was supported by funding from ExxonMobil to the UTeach Institute and by the National Math and Science Initiative. The project was instigated by questions posed by the Howard Hughes Medical Institute. Thanks are due to Celeste Alexander and Cindy Corn for frequent technical assistance in accessing Texas Educational Research Center data. Ben David carried out the work to separate out students taking modified examinations. Paul Hanselman carefully read a draft manuscript and provided many helpful suggestions. Thanks as well for suggestions to Rich Brown of the National Math and Science Inititiative, Amy Winters, Kim Hughes and Pamela Romero of the UTeach Institute, and Valerie Ringland.

References

114th Congress (2015). Every student succeeds act. Available from: `https://www.congress.gov/bill/114th-congress/senate-bill/1177/text` [Last accessed October 2016].

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, 25(1):95–135.

Barth, P., Dillon, N., Hull, J., & Higgins, B. H. (2016). Fixing the holes in the teacher pipeline: An overview of teacher shortages. Available from: `http://www.centerforpubliceducation.org/Main-Menu/Staffingstudents/An-Overview-of-Teacher-Shortages-At-a-Glance` [Last accessed November, 2016].

Bolker, B., Bates, D., Maechler, M., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Grothendieck, G., & Green, P. (2016). Package lme4. Available from: `https://cran.r-project.org/web/packages/lme4/lme4.pdf` [Last accessed November, 2016].

Bonner, L. (February 3, 2016). Enrollment plunges at UNC teacher prep programs. Charlotte Observer. Available from: `http://www.charlotteobserver.com/news/local/education/article58311793.html` [Last accessed November 2016].

Boyd, D., Goldhaber, D., Lankford, H., & Wyckoff, J. (2007). The effect of certification and preparation on teacher quality. *Future of Children*, 17(1):45 – 68.

Boyd, D., Grossman, P., Hammerness, K., Lankford, H., Loeb, S., Ronfeldt, M., & Wyckoff, J. (2012). Recruiting effective math teachers: Evidence from new york city. *American Educational Research Journal*, 49(6):1008 – 1047.

Boyd, D., Grossman, P., Lankford, H., Loeb, S., & Wykoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis*, 31:416–440.

Clark, M. A., Chiang, H. S., Silva, T., McConnell, S., Puma, K. S. A. E. M., Warner, E., & Schmidt, S. (2013). The effectiveness of secondary math teachers from Teach For America and the Teaching Fellows Programs. Institute for Education Sciences. Available from: `http://ies.ed.gov/ncee/pubs/20134015/` [Last accessed August 2014].

College Board (2016). AP program participation and performance data 2016. Available from: `https://research.collegeboard.org/programs/ap/data/participation/ap-2016` [Last accessed October 1, 2016].

Constantine, J., Player, D., Silva, T., Hallgren, K., Grider, M., & Deke, J. (2009). An evaluation of teachers trained through different routes to certification. final report. *National Center for Education Evaluation and Regional Assistance*.

Cowan, J. & Goldhaber, D. (2016). National Board Certification and Teacher Effectiveness: Evidence From Washington State. *Journal of Research on Educational Effectiveness*, 9(3):233–258.

Darling-Hammond, L., Holtzman, D. J., Gatlin, S. J., & Heilig, J. V. (2005). Does teacher preparation matter? Evidence about teacher certification, Teach for America, and teacher effectiveness. *Education Policy Analysis Archives*, 13(42).

Decker, P. T., Mayer, D. P., & Glazerman, S. (2004). The effects of Teach For America on students: Findings from a national evaluation. Available from: `https://www.mathematica-mpr.com/-/media/publications/pdfs/teach.pdf` [Last accessed October, 2016].

Duncan, A. (2010). Teacher preparation: Reforming the uncertain profession. Available from: `http://www.ed.gov/news/speeches/teacher-preparation-reforming-uncertain-profession` [Last accessed June 2016].

Gates Foundation (2012). Gathering feedback for teaching: Combining high quality observations with student surveys and achievement gains. Research paper. Available from: `http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf` [Last accessed 9 December 2012].

Goldhaber, D. & Chaplin, D. D. (2015). Assessing the 'Rothstein falsification test': Does it really show teacher value-added models are biased? *Journal of Research on Educational Effectiveness*, 8:8–34.

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.

Gordon, R., Kane, T. J., & Staiger, D. O. (2006). Identifying effective teachers using performance on the job. Available from: `www.brookings.edu/views/papers/200604hamilton_1.pdf` [Last accessed November 2010].

Greenberg, J., McKee, A., & Walsh, K. (2013). Teacher prep review: A review of the nation's teacher preparation programs. Available from: `http://www.nctq.org/dmsView/Teacher_Prep_Review_2013_Report`.

Grimmett, P. P. & Young, J. C. (2012). *Teacher Certification and the Professional Status of Teaching in North America: The new battleground for public education.* Information Age Publishing, Charlotte, NC.

Grossman, P., Loeb, S., & Harvard University, G. S. o. E. (2008). *Alternative Routes to Teaching: Mapping the New Landscape of Teacher Education.* Harvard Education Press.

Guarino, C. M., Santibanez, L., & Daley, G. A. (2006). Teacher recruitment and retention: a review of the recent empirical literature. *Review of Educational Research*, 76(2):173–208.

Guzdial, M. (2012). Why isn't there more computer science in US high schools? Available from: `http://cacm.acm.org/blogs/blog-cacm/156531-why-isnt-there-more-computer-science-in-u-s-high-schools/fulltext` [Last accessed Retrieved October 2016.].

Hanushek, E. (1971). Teacher characteristics and gains in student achievement: Estimation using micro data. *American Economic Review*, 61:280–288.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review*, 30(3):466–479.

Harris, D. N. & Sass, T. R. (2011a). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, 95(7-8):798–812.

Harris, D. N. & Sass, T. R. (2011b). Teacher training, teacher quality and student achievement. *Journal of Public Economics*, 95(7-8):798–812.

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

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

Hess, F. M. (2001). Tear down this wall: The case for a radical overhaul of teacher certification. Available from: `http://www.ppionline.org/documents/teacher_certification.pdf` [Last accessed October 2010.].

Kamnetz, A. (2014). For teachers many paths into the classroom...some say too many. Available from: `http://www.npr.org/sections/ed/2014/09/12/347375798/for-teachers-many-paths-into-the-classroom-some-say-too-many` [Last accessed Accessed October 1, 2016].

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.

Kane, T. J. & Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Available from: `http://www.nber.org/papers/w14607` [Last accessed November 2010].

Kinsler, J. (2012). Assessing Rothstein's critique of teacher value-added models. *Quantitative Economics*, 3(2):333–362.

Koedel, C. & Betts, J. R. (2011). Does student sorting invalidate Value-Added Models of teacher effectiveness? an extended analysis of the Rothstein critique. *Education Finance and Policy*, 6(1):18–42.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, 47:180–195.

Koedel, C., Parsons, E., Podgursky, M., & Ehlert, M. (2012). Teacher preparation programs and teacher quality: Are there real differences across programs? CALDER Working paper 79. Accessed August, 2016. Available from: `www.caldercenter.org/sites/default/files/WP-79.pdf`.

Laczko-Kerr, I. & Berliner, D. (2002). The effectiveness of "teach for america" and other under-certified teachers. *Education Policy Analysis Archives10*, 10. Available from: `http://epaa.asu.edu/ojs/article/view/316`.

Mellor, L., Lummus-Robinson, M., Brinson, V., & Dougherty, C. (2008). Linking teacher preparation programs to student achievement in texas. *Unpublished*. As the main conclusion of this study was that teacher preparation pathways could not be detected in the dataset it was not easy to publish.

Morgan, S. & Winship, C. (2015). *Counterfactuals and Causal Inference: Methods and Principles for Social Research. Second Edition.* Cambridge University Press, second edition.

National Commission on Excellence in Education (1983). A nation at risk. Available from: `http://www2.ed.gov/pubs/NatAtRisk/index.html` [Last accessed October 2016].

National Science Foundation (2016). Robert noyce teacher scholarship program. Available from: `http://www.nsf.gov/pubs/2016/nsf16559/nsf16559.pdf` [Last accessed October 2016].

NRC (2010). *Preparing Teachers: Building Evidence for Sound Policy.* National Research Council. Committee on the Study of Teacher Preparation Programs in the United States.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, 26:237–257.

Pearl, J. (2009). *Causality.* Cambridge University Press, second edition.

Rivkin, S., Hanushek, E., , & Kain, J. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73:417–458.

Rothstein, J. (2009). Student sorting and bias in value added estimation: selection on observables and unobservables. *Education Finance and Policy*, 4:537–571. Available from: http://eml.berkeley.edu/~jrothst/publications/rothstein_VAM_EFP.pdf.

Sanders, W. & Rivers, J. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. University of Tennessee Value-Added Research and Assessment Center, Knoxville.

Sass, T. R. (2011). Certification requirements and teacher quality: A comparison of alternative routes to teaching. working paper 64. *National Center for Analysis of Longitudinal Data in Education Research.*

Schools and Staffing Survey (2012). Education and certification qualifications of departmentalized public high-school level teachers. Table 2.

Sentell, W. (2016a). BESE narrowly endorses overhaul in teacher training. The Advocate. Available from: http://www.theadvocate.com/baton_rouge/news/education/article_727ec7f6-8fb5-11e6-8d6e-3b3e8d39152e.html [Last accessed October 2016].

Sentell, W. (2016b). Bid to overhaul teacher training sparks new arguments. The Advocate. Available from: http://www.theadvocate.com/baton_rouge/news/education/article_ceb82166-7f6b-11e6-b777-7f901ee69b3a.html [Last accessed October 2016].

Staiger, D. O. & Rockoff, J. E. (2010). Searching for Effective Teachers with Imperfect Information. *Journal of Economic Perspectives*, 24(3):97–117.

Suell, J. L. & Piotrowski, C. (2007). Alternative teacher education programs: a review of the literature and outcome studies. *Journal of Instructional Psychology*, 34(1):54–58.

Teachers of Tomorrow (2016). Florida Teachers of Tomorrow receives state approval. Available from: https://www.edreform.com/2016/08/florida-teachers-of-tomorrow-receives-state-approval/ [Last accessed October 1, 2016].

Turner, H. M., Goodman, D., Adachi, E., Brite, J., & Decker, L. E. (2012). Evaluation of Teach For America in Texas schools. Edvance. Available from: http://edvanceresearch.com/wp-content/uploads/2015/06/Evaluation-of-Teach-For-America-in-Texas-Schools.pdf.

US Department of Education (2016a). Teacher preparation issues: Final regulations. Available from: http://www2.ed.gov/documents/teaching/teacher-prep-final-regs.pdf [Last accessed October 2016].

US Department of Education (2016b). Title II Higher Education Act Data Tools. Available from: https://title2.ed.gov/Public/DataTools/Tables.aspx [Last accessed October 2016].

von Hippel, P. T., Osborne, C., Lincove, J., Bellow, L., & Mills, N. (2014). The challenges of seeking exceptional teacher preparation programs among many noisy estimates. *SSRN working paper 2506935*. Available from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2506935 [Last accessed October 1, 2016].

Wayne, A. & Youngs, P. (2003). Teacher characteristics and student achievement gains: A review. *Review of Educational Research*, 73(1):89–122.

List of Tables

List of Figures