Journal of Public Economics 100 (2013) 61–78

Contents lists available at SciVerse ScienceDirect

Journal of Public Economics

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

The dynamics of teacher quality ☆

Matthew Wiswall

Arizona State University, W.P. Carey School of Business, Department of Economics, United States

ARTICLE INFO

Article history:
Received 19 January 2011
Received in revised form 21 January 2013
Accepted 26 January 2013
Available online 8 February 2013

JEL classification:
I2
J4

Keywords:
Economics of education
Teacher quality
Teacher dynamics
Teacher experience
Fixed effect estimators

ABSTRACT

An extensive literature finds that while teachers vary considerably in initial quality there are limited teacher quality dynamics: except for the first few years of teaching, teacher quality does not improve over the course of a teacher's career. This study evaluates the importance of various modeling restrictions to the key findings of this literature. Using data covering all 5th grade public school teachers from the state of North Carolina, I replicate the findings of the previous literature using their restrictive experience assumptions. However, using an unrestricted experience model I find that for mathematics achievement there are high returns to later career teaching experience, about twice as much dispersion in initial teacher quality as previously estimated, and a pattern of negative selection where high quality teachers are more likely to exit.

© 2013 Elsevier B.V. All rights reserved.

> *"Once somebody has taught for three years their teaching quality does not change thereafter."* Bill Gates, 2009[1]

> *"We know that experience makes a difference in student achievement — teachers get better."*Bill Raabe, National Education Association, 2010[2]

1. Introduction

How does teacher quality evolve over a teacher's career? The answer to this question affects several areas of the current policy debate about reforming the hiring, retention, and compensation policies for public school teachers. If teachers become more effective through years of accumulated experience and their quality increases over their career, then teacher turnover should be a major concern and schools should base compensation and promotion on rewarding experienced teachers and encouraging retention. If however teacher quality does not increase with experience, and instead teacher quality is largely constant over a teacher's career, then public schools should not base compensation and promotion as rigidly on teaching experience and should instead focus on hiring and retaining teachers with the highest level of innate quality.

With the availability of large administrative data sets containing repeated observations of test score outcomes for students and teachers, the literature on teacher quality has made important advances in understanding the determinants of quality teaching.[3] The current consensus in this literature is that while there is dispersion in quality among teachers, the return to experience in teaching is

☆ I thank two anonymous referees and Mark Duggan, Co-Editor, for their detailed and helpful feedback. I thank Stephane Bonhomme, Sean Corcoran, Josh Kinsler, and Amy Schwartz for comments on older versions of the paper. I also thank Jesse Rothstein for sharing data preparation code for the North Carolina data. I also thank the North Carolina Education Data Research Center at Duke University for assembling, cleaning, and making available the confidential data used in this study.

E-mail address: matt.wiswall@gmail.com.

[1] Speech from Feb. 2009, ted.com website. Full quote: "What are the characteristics of this top quartile? What do they look like? You might think these must be very senior teachers. And the answer is no. Once somebody has taught for three years their teaching quality does not change thereafter." I thank Sean Corcoran for pointing out the quotation.

[2] Quoted in *New York Times* by Sam Dillon Published, November 19, 2010, "Gates Urges School Budget Overhauls."

[3] This research continues to expand, with over 400 articles already completed (Hanushek and Rivkin, 2006). See Hanushek and Rivkin (2006) and Wayne and Youngs (2003) for summaries and reviews of the more recent literature and how the older literature motivates this research. Examples of the most recent research include: Harris and Sass (2011) using data for Florida; Clotfelter et al. (2006b, 2007) using data for North Carolina; Hanushek et al. (2005b) using data for Texas; Rockoff (2004) using data for New Jersey; and Aaronson et al. (2007) using data for Chicago; Nye et al. (2004) using data for Tennesee; and Kane et al. (2008) and Boyd et al. (2005) using data for New York City.

0047-2727/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jpubeco.2013.01.006

small, except for the first few years of teaching.[4] Teachers have some gains in effectiveness in the first few years after entry, but later teaching experience after the initial years contributes little or nothing to a teacher's quality or effectiveness in the classroom. The dynamics of teacher quality are therefore believed to be quite limited. These findings have led researchers to conclude that public schools in the US should abandon the reliance on experience based salary and promotion systems and instead use measures of teacher quality to selectively retain those early career teachers with the highest measured quality.

In this paper, I find that low estimates of later teaching experience are generated by overly restrictive empirical specifications, as have been used in previous studies. Using matched student–teacher data for public schools in the state of North Carolina and a large sample of students and 5th grade teachers, I replicate the earlier findings of small returns to later teacher experience using the previous restricted models. However, when I relax these restrictions, and allow a flexible non-parametric relationship between experience and teacher quality, combined with a teacher fixed effect model that uses within teacher variation in experience to identify the return to experience, I find that teaching experience has a substantial and statistically significant impact on mathematics achievement, even beyond the first few years of teaching. I estimate that a teacher with 30 years of experience has over 1 standard deviation higher measured quality than new, inexperience teachers, and about 0.75 standard deviations higher measured mathematics effectiveness than a teacher with 5 years of experience. In comparison, estimates on the same data sample using previously restricted models suggest that experienced teachers have between 0.1 and 0.2 standard deviations higher quality than new teachers, with almost all of these gains in the first few years of teaching. In contrast to the mathematics results, in my unrestricted models, I find no statistically significant relationship between later career teaching experience and student reading scores. This stark difference in the experience returns for mathematics and reading is hidden by the restrictive models previously estimated in which the estimated experience returns are similar across subjects.

While considerable attention has been paid to the specification of the student learning process, in particular the dynamics of student learning and the rate at which student skills accumulates or depreciates over time (e.g. Todd and Wolpin, 2003), less attention has been paid to the analogous process of teacher learning or teacher quality dynamics. The reason I obtain such different results is that I adopt a more flexible and robust specification of the evolution of teacher quality.

There are two key dimensions of modeling teacher quality. The first modeling dimension comprise the restrictions imposed on the relationship between accumulated teaching experience and unobserved determinants of teacher quality. As is well known, there is considerable attrition from teaching, with some teachers leaving after a few years and others staying much longer (see Fig. 4). When teachers exit teaching endogenously, OLS or random effects estimates (RE) of the return to teaching experience are biased since the level of experience is correlated with unobserved components of teacher quality. If teacher exits are based on a career constant teacher quality component, teacher fixed effects (FE) models can provide consistent estimates of the return to teaching estimates.

The second dimension of modeling teacher quality is the restrictions imposed on the relationship between teaching experience and quality. Many of the specification used in previous studies allow for high returns to early years of experience but severely restrict the later teaching experience dynamics. I discuss a number of these specifications in turn and investigate the importance of these restrictions by estimating each of these specifications on the NC data sample. While there is some variation in results, these restricted specifications all provide much lower estimates of later teaching experience and lower estimates of teacher quality dispersion. Remarkably, these restricted specifications produce experience return estimates that are quite similar across the different teacher quality models: OLS, RE, and FE models. However, when I relax the experience restrictions using a non-parametric specification, I find very different results across the models. I also show that some simple smooth parametric in experience models, even a one parameter linear in experience model, produce estimates of the return to experience that approximate the non-parametric estimates much more closely than any of the previously used experience models.

The key issue introduced in this paper is that the teacher fixed effects estimator using the particular restricted dummy variable specifications from the current literature implicitly use only selected sub-samples of teachers and career years in the estimation. If there is no variation in experience within the particular window of observations that are observed in the data *according to the particular experience model*, then these teachers do not contribute at all to the FE estimates of teaching experience. I find that for one such restricted specification used in the previous literature, the teacher FE estimator computed on the NC sample uses only 1/3 of the total sample of teachers and ignores the 2/3 of teachers who have no variation in experience according to the restricted experience specification. Because the excluded teachers are primarily more experienced, this implicit selection of the FE estimator produces estimates that do not represent some average of experience returns for teachers of all experience levels. Importantly, this issue is specific to the teacher FE estimator and is avoided by pooled OLS (teacher characteristics model) estimators or teacher RE estimators. However, the pooled OLS and RE estimators are biased due to selected exits from teaching, as discussed above. The two separate sources of bias cause the pooled OLS and RE estimates of experience returns to resemble those obtained using the restricted FE estimated, although all of these estimates are very different from those obtained using the unrestricted, non-parametric in experience, teacher FE estimator.[5]

The plan for the remainder of the paper is as follows. The next section discusses the previous literature. Section 3 covers model specification and identification issues related to the teacher quality contribution to education production. The following sections present the data and estimation framework. I then present results on the return to experience, dispersion of teacher quality, and dynamic teacher sorting. I conclude with a discussion and interpretation of the results.

2. Education production function models

2.1. Outcomes, students, and teachers

I consider a general setup in which students indexed *i* are taught by teachers indexed *j*. To avoid additional notation for grade levels, calendar periods, and schools, I refer to a generic matching of student *i* to teacher *j*. As detailed in the estimation sections below, I estimate the model for 5th grade teachers and include unrestricted fixed effects for calendar periods and schools to absorb their contribution. I consider

[4] Harris and Sass (2011) provide a comprehensive recent summary of the findings on effectiveness of pre-service education, in-service training, and experience. Several recent papers have found some limited gains in effectiveness from formal teacher in-service training. Generally most prior research finds gains in effectiveness in primarily in the first few years of teaching and limited gains thereafter. A few papers find some evidence of gains to later accumulated experience, including Clotfelter et al. (2007) using data from North Carolina, Harris and Sass (2011) using data from Florida, and Papay and Kraft (2010). We replicate the specifications of these papers and others using the North Carolina data and find substantially higher returns to experience in mathematics using unrestricted, non-parametric specifications than using the prior specifications.

[5] This evidence for teachers is consistent with studies examining wage growth patterns across all occupations, in which there are typically positive and large returns to experience, even allowing for non-random attrition from jobs (e.g. Topel, 1991; Altonji and Williams, 1998). But wages are not always directly tied to current productivity (see Flabbi and Ichino, 2001). Other research has found that returns to experience are an important feature of labor supply decisions, especially for women (e.g. Olivetti, 2006).

the general case where teachers teach for potentially more than one period. t denotes the period in teacher j's career she is matched to student i, with the first period of her career (when she has no previous experience) denoted $t=0$. Student outcome y_{ijt} (in our data, an end of grade test score in mathematics or reading) is the outcome for student i who is taught by teacher j in period t of the teacher's career.[6]

Following the existing literature, outcome y_{ijt} is decomposed into an additively separable student i component α_i and a teacher j by career year t component π_{jt}:

$$y_{ijt} = \alpha_i + \pi_{jt} + \epsilon_{ijt}, \tag{1}$$

where ε_{ijt} is a mean zero residual term. I focus on identifying and estimating the π_{jt} term. I therefore leave the specification of the student component α_i quite general at this point and follow the current literature in exploring different specifications of it in the empirical analysis, including incorporating lagged student test scores from previous grades and student level fixed effects. Note that α_i represents the student's contribution to the outcome when matched to teacher j at career year t; this term is not necessarily assumed to be constant across grade levels.[7] Note that t indexes the year of the teacher's career, not the grade level or calendar period.

2.2. Teacher quality models

I next discuss the array of teacher quality models used in the current literature. These models take the form:

$$\pi_{jt} = \theta_j + f\left(x_{jt},\beta\right) + \phi_{jt}, \tag{2}$$

where teacher quality $q_{jt} \equiv \theta_j + f(x_{jt},\beta)$ has two components: θ_j is the time invariant teacher quality component for teacher j and $f(x_{jt},\beta)$ is some specified function of the level of experience $x_{jt} \in \{0,1,2,\ldots,T\}$ and a finite dimensional parameter vector β. ϕ_{jt} is the remaining residual classroom variation in quality. q_{jt} reflects the heterogeneity in teacher quality across different teachers j and across the years of the teacher's career t.[8] Dynamics in teacher quality arise with $q_{jt} \neq q_{jt'}$ for $t \neq t'$. One special case is if there are no career dynamics then the level of quality remains the same across the teacher career years: $q_{jt} = q_{j0}$ for all t.

I consider restrictions on this model of two types: i) restrictions on the degree of sorting by unobservable components (i.e. the choice between a teacher characteristics (OLS), teacher random effects (RE), or teacher fixed effects (FE) models), and ii)restrictions on the functional form for experience $f(x_{jt},\beta)$ (i.e. whether to impose some particular parametric restriction, such as a linear restriction $f(x_{jt},\beta) = \beta x_{jt}$ or to allow for a more unrestricted relationship). I discuss both types of modeling restrictions and show that both are critical to identifying teacher quality heterogeneity and dynamics.

2.2.1. Teacher characteristics models (OLS)

The first model I consider is a *teacher characteristics* model:

$$\pi_{jt} = W'_j\Gamma + f\left(x_{jt},\beta\right) + \eta_{jt}, \tag{3}$$

where W_j is a vector of time invariant observable teacher characteristics (e.g. undergraduate education or licensure examination score), including an intercept. η_{jt} is a mean zero residual teacher quality component, which I can express, without loss of generality, as a function of the unobserved time invariant teacher component and a residual component: $\eta_{jt} = \theta^*_j + \phi_{jt}$ and $\theta_j = \theta^*_j + W'_j\Gamma$. This expression makes clear that the time invariant teacher component θ_j consists of the contribution from observable characteristics $W'_j\Gamma$ and unobservable components θ^*_j.

The major distinguishing feature of the teacher characteristics model is that the time invariant teacher quality component is assumed to be determined by a set of time invariant observable teacher characteristics W_j. A special case of this model is where W_j includes only an intercept, $W_j = 1$ for all j. Teacher characteristics models can be estimated using a pooled OLS estimator. Teacher characteristics models encompass both models that are estimated using a single cross-section of classrooms (e.g. Chetty et al., 2010), and approaches that pool multiple years of data but do not include unobserved teacher random or fixed effects (e.g. Clotfelter et al., 2007).

In the teacher characteristics model, bias in the estimation of the return to experience β occurs when unobserved components of teacher quality, given by θ^*_j, are correlated with accumulated experience x_{jt}. An example is if θ^*_j reflects a teacher's level of general skills, skills both productive in teaching and in outside occupations, and teachers with higher levels of general skills exit teaching because they receive higher wage offers in non-teaching occupations. This *teacher self-selection* is a separate source of endogeneity, distinct from non-random student to teacher matching, which is the main focus in much of the teacher quality literature. Even in the extreme case where students and teachers are randomly matched, teacher self-selection implies that the OLS estimator of Eq. (3) still would be a biased estimator of β as x_{jt} would be correlated with $\eta_{jt} = \theta^*_j + \phi_{jt}$.

2.2.2. Teacher random effects (RE) models

A second type of teacher quality model allows for a distribution of unobserved teacher effects but assumes a restricted relationship between these effects and teaching experience. The random effects (RE) model assumes a particular θ_j distribution, and assumes θ_j is independent of both teaching experience and the residual component ϕ_{jt}. In addition, this model assumes that the residual time varying component is uncorrelated with accumulated experience. The RE teacher model potentially suffers from the same teacher self-selection issue as the teacher characteristics model. With teachers self-selecting when to exit from teaching, at least partially based on their level of teacher quality θ_j, x_{jt} is no longer independent from θ_j.[9]

2.2.3. Teacher fixed effects (FE) models

As a response to the limitations of the teacher characteristics and RE models, another set of models uses teacher fixed effects (FE). These types of models allow for an unrestricted relationship between the time invariant teacher quality component and the level of experience. In the FE model, teacher self-selection based on θ_j is allowed. However, the model produces biased estimates if accumulated teaching experience is correlated with time varying shocks ϕ_{jt}. An example of this source of bias is if a particularly unruly class of students assigned to the teacher in period t (a negative ϕ_{jt} realization) causes the teacher to leave teaching. This type of behavior creates a correlation between ϕ_{jt} and x_{jt}. This kind of *dynamic teacher selection* is the teacher career analog of the dynamic matching of teachers to students as when students are assigned to teachers based on student or teacher performance in the previous grade level (Rothstein, 2010).

[6] This framework could incorporate a gains model in which y_{ijt} is the difference in some outcome between two different periods or grade levels. As discussed in the estimation section, I incorporate dynamics in student learning using a lagged score model with variables for previous grade level test scores.

[7] For simplicity of exposition, I ignore student–teacher match specific components here. However, in the empirical application, I allow for these types of match specific components by including variables for whether the student and teacher are of the same gender or race. See Ouazad (2008) and Lockwood and McCaffrey (2009) for a recent discussion of student match specific gains in teaching.

[8] In our empirical application, I discuss the importance of distinguishing the teacher quality component from the effect of other classroom characteristics, such as class size and student composition. The data includes a range of classroom characteristics.

[9] In more elaborate models, the RE model is specified as a multilevel model where the distribution of the random effect depends on observed teacher characteristics (e.g. Nye et al., 2004). In this case, unbiased estimates of teacher dynamics rely on the correct specification of the distribution of teacher effects.

2.3. Teaching experience functional form restrictions

I next discuss the restrictions on the teacher quality dynamics embodied in the restrictions on the $f(x_{jt},\beta)$ experience function. Much of the current literature restricts this function using a series of selected dummy variables for different experience levels, with greater saturation for lower levels of experience to capture the possibility that there are higher returns to early career experience. What has not previously been recognized is that these different dummy variable constructions and experience restrictions interact differently with the particular estimating framework (OLS, RE, or FE). Specifically, as I show here, estimates of teacher FE models with commonly used restricted experience specifications implicitly use selected sub-samples of teachers and lead to estimates of the effect of experience that reflect only a limited range of teacher career years. As I show in the empirical analysis, these restricted models can cause severe biases in the estimated return to teaching experience, the distribution of teacher quality, and the patterns of dynamic teacher sorting.

To motivate this issue, consider an example using the matched student-teacher North Carolina data set (discussed in detail below). To illustrate the issue, I examine the following experience specification from Kane et al. (2008).[10] The Kane et al. (2008) experience specification is

$$f\left(x_{jt},\beta\right)=\beta_1 1\left\{x_{jt}=1\right\}+\beta_2 1\left\{x_{jt}=2\right\}+\beta_3 1\left\{x_{jt}=3\right\}+\beta_{4p} 1\left\{x_{jt}\geq 4\right\} \quad (4)$$

where $1\{\cdot\}$ are dummy/indicator variables. This specification includes dummy variables for the first 3 years of experience and one additional dummy variable that equals 1 for teachers with 4 or more years of experience. The omitted category is new teachers, with 0 years of experience. β_1,β_2,β_3 are intended to capture returns to experience for the first 3 years of experience and β_{4p} is intended to capture the return to higher levels of experience. The specification is restricted since it assumes that the 5th and higher years of experience provide *no* additional increase in teacher quality: $f(x+1)-f(x)=0$ for all $x\geq 5$. While the specification allows experienced teachers to have a higher *level* of teacher quality than teachers with 3 years of experience, with that difference in quality given by a constant $f(x)-f(3)=\beta_{4p}-\beta_3$ for $x\geq 4$, the functional form restricts the *marginal* change in teacher quality after 4 years of experience to be 0. In contrast, an alternative linear in experience model, $f(x)=\beta x$, allows for some marginal change in teacher quality in all years, although that marginal effect of experience is restricted in this linear model to be a constant: $f(x+1)-f(x)=\beta$ for all $x\geq 0$.

Next I demonstrate that this type of restricted experience specification only captures returns to experience for a selected sub-sample of teachers who are observed in the first years of their career. Note that most matched teacher-student data consists of unbalanced panel data with censoring from below and above: for younger teachers, I observe only the first years of their career and do not observe later, unrealized, future years; for older teachers, the data do not include early career observations as the data collection begins after they have begun their career. Using the Kane et al. (2008) specification (4) for teachers who are observed only in the years of their career after accumulating 4 years of experience, there is no variation in *any* of the included experience variables. The variables $1\{x_{jt}=k\}=0$ for $k=1,2,3$ and $1\{x_{jt}\geq 4\}=1$ for all t career years observed for these experienced teachers. In the standard terminology of panel data analysis, these teachers are permanent "stayers", and hence contribute nothing to the estimated β experience returns, including the FE estimator of β_{4p}, which purportedly captures the experience returns of older teachers. Only relatively inexperienced teachers are "switchers" in this specification and contribute to the estimates of $\beta_1,\beta_2,\beta_3,\beta_{4p}$.

A simple way to see this is to compare columns (1) and (2) of Table 1. The first column uses the full sample of 5th grade teachers in the NC data, which range from having 0 years of experience (new teachers) to 40 years of experience. The second column uses a restricted sample of teachers who are observed in the first 5 years of their career (with experience ranging from 0 to 4 years) and excludes teachers who are first observed in the 6th or later year of their career (with experience ranging from 5 to 40 years). The restricted sub-sample of young teachers in column (2) is approximately 1/3 of the full sample. The key result is that the difference in sample selection does not affect the teacher FE estimates of the experience coefficients at all. This is because the 2/3 of teachers excluded from this sample in column (2) do not contribute to the estimates of the restricted experience model. While this is reasonable for the dummy variables capturing the first 3 years of experience returns $\beta_1,\ldots,\beta_3$, the remarkable aspect of Table 1 is that the estimate of β_{4p} is similarly unaffected. β_{4p} does not represent some average of the effect of experience for older teachers with between 5 and 40 years of experience. Instead it represents an estimate of experience for relatively inexperienced teachers, those with experience around 5 years for which we have some early career observations, and does not reflect at all the sample of more experienced teachers.[11]

The Appendix provides an analysis of the FE estimator in the case of these type of restricted experience specifications. As discussed in the Appendix, an important difference between the FE and pooled OLS estimators in these cases is that while the pooled OLS estimator would always use the full sample in the estimator, the FE estimator uses only the sample with at least some variation in the time varying experience variables. The classification of the sub-sample that are "stayers" and "switchers" depends on the restrictions imposed on the $f(x_{jt},\beta)$ experience specification. Therefore, as studies change their specification of $f(x_{jt},\beta)$ they implicitly use different sub-samples of teachers in the estimation.[12]

In the teacher quality literature, there are a number of other restricted experience specifications. Clotfelter et al. (2007) uses a number of dummy variables for experience intervals:

$$f\left(x_{jt},\beta\right)=\beta_{1-2}1\left\{1\leq x_{jt}\leq 2\right\}+\beta_{3-5}1\left\{3\leq x_{jt}\leq 5\right\}+\beta_{6-12}1\left\{6\leq x_{jt}\leq 12\right\}+\beta_{21-27}1\left\{21\leq x_{jt}\leq 27\right\}+\beta_{28p}1\left\{x_{jt}\geq 28\right\} \quad (5)$$

This specification restricts the marginal effect of additional experience to be 0 at several levels. For example, while the marginal effect of initial experience is $f(1)-f(0)=\beta_{1-2}$, the marginal effect of an additional year is $f(2)-f(1)=0$. Similarly, higher levels of experience are restricted to have a zero marginal effect: any experience accumulated between 3 and 5 years, 6 and 12 years, and so on. Harris and Sass (2011) use a similar specification with dummy variable intervals at 1–2, 3–4, 5–9, 10–14, 15–24, and 25 plus. Rockoff (2004) uses a specification which is cubic in experience through the first 10 years, but imposes a zero marginal return to experience after the first 10 years:

$$f\left(x_{jt},\beta\right)=\left(\beta_1 x_{jt}+\beta_2 x_{jt}^2+\beta_3 x_{jt}^3\right)1\left\{x_{jt}<10\right\}+\beta_{10p}1\left\{x_{jt}\geq 10\right\} \quad (6)$$

[10] Kane et al. (2008) estimate this model using New York City student-teacher matched data. Since the focus of Kane et al. (2008) is differential returns to experience for alternative versus regular licensed teachers, their specification includes a number of other variables, which I do not include here. As I discuss below, including school effects and other variables in the full specification, the estimates of this model using NC data and those using NYC data are quite similar.

[11] It is of course possible that the difference in the estimates is consistent with a hypothesis of no return to later career experience. However, as I show below, estimates using other FE estimators provide evidence of positive returns to later career experience.

[12] See Loken et al. (2012) for a discussion of the particular weighting of the linear FE estimator in another context.

Table 1
An example: experience restrictions in a teacher fixed effect analysis using unbalanced panel data.

	Full sample	Teachers observed with experience <5 years
Exper. 1 Year	.0816	.0816
Exper. 2 Years	.113	.113
Exper. 3 Years	.136	.136
Exper. 4+ Years	.148	.148
Observations	11,460	3465

Notes: This table shows that using a restricted teacher experience specification in a teacher FE analysis implicitly uses only a subset of teacher observations. This table uses the experience specification of Kane et al. (2008). The experience coefficient estimates in columns (1) and (2) are identical, even though column (1) uses the full sample of teachers and column (2) uses only the sample of teachers observed with 4 or fewer years of experience during the sample period and drops those teachers observed during the sample period with 5 or more years of experience. Standard errors are not reported here. These specifications do not include any other control variables. See below for full results including the full set of control variables. This table is illustrative of the econometric issue only. The omitted category is teachers with 0 years of experience (new teachers in their first year of teaching).
Source: North Carolina matched student-teacher data.

Hanushek et al. (2005b) specifies teacher experience similarly to Kane et al. (2008), but uses experienced teachers as the omitted reference category:

$$f\left(x_{jt},\beta\right) = \beta_0 1\left\{x_{jt}=0\right\} + \beta_1 1\left\{x_{jt}=1\right\} + \beta_2 1\left\{x_{jt}=2\right\} + \beta_3 1\left\{x_{jt}=3\right\} + \beta_4 1\left\{x_{jt}=4\right\} \quad (7)$$

The omitted category is teachers with 5 or more years of teaching experience. The marginal return to experience, relative to new teachers, is given by $f(x)-f(0)=\beta_x-\beta_0$ for $x=1,2,3,4$. For higher levels of experience, the marginal return to experience is given by $f(x)-f(0)=-\beta_0$ for $x\geq 5$. Like the other specifications, the Hanushek et al. (2005b) specification imposes a restriction on the marginal return to later career experience. The marginal return to later career experience is 0 since $f(x)-f(x-1)=0$ for $x\geq 5$.

3. Data

This section describes the data and sample selection criteria. In general, I follow procedures used by previous researchers. As evidence that our sample selection is similar to others, I show that the baseline results, where I replicate existing models, are similar to existing findings.

I use restricted use matched student to teacher administrative data from the state of North Carolina. The data includes nearly all teachers and students at North Carolina public schools who were present in the NC public schools from 1996 to 2005. The teacher of record for each student is the teacher proctoring the end of grade examination.[13] At the student level, the data includes end of grade test scores in reading and mathematics, along with basic demographic data for the students (gender, race, and free lunch eligibility indicators of family income). Reading and mathematics scores for each grade and year are normed (prior to sample selection) so that they have 0 mean and standard deviation 1 in the population. At the teacher level, the data collection includes personnel records for all teachers, including demographic information (race and gender), records of education (undergraduate and graduate degrees), licensure status and initial licensure examination scores, and years of experience (in the NC public schools). Years of experience is measured using payroll records for the state of NC and excludes any experience that a teacher may have accumulated in other schools, including private schools in NC or public or private schools in other states. The teacher licensing score is generated by averaging scores on the examinations teachers took prior to licensure. Each of these scores is normed by using the full set of recorded tests taken in that testing year so that each test has mean 0 and standard deviation 1 in the year taken.

3.1. Sample selection

I focus on a sub-set of the original data comprising 5th grade teachers but make use of 3rd and 4th grade data as well. I focus on the 5th grade teachers for two reasons. First, this is one of the last grades in which nearly all students have the same teacher for major subject areas. Focusing on 5th grade teachers then provides a more interpretable match between end of grade test scores and teachers than in higher grades where students may have multiple teachers for separate subjects. Second, for students entering 5th grade, I have available past test scores in reading and mathematics for 3rd and 4th grades. The availability of this prior test score information provides the maximum amount of information to control for student's prior level of ability entering 5th grade and thereby minimize issues of non-random sorting of students to teachers.

Our initial sample of students excludes students missing demographic information, special education students, students who repeat elementary grades, and students without complete records of mathematics and reading exam scores for 3rd, 4th, and 5th grades (thereby excluding transfer students or students who enter the NC system after 3rd grade or left before 5th grade). Our initial sample of teachers excludes teachers missing demographic information, teachers who taught exclusively special education or honors/gifted students, or teachers who did not teach both mathematics and reading in the same classroom. I further exclude classrooms (and therefore the teacher and student observations for that grade year) with unusually small class sizes (less than 10 students), unusually large class sizes (greater than 50 students), and all classrooms that had students from multiple grades. I include class size as a control variable in all specifications.

After selecting an initial sample of students and teachers, I then impose additional sample selection criteria to ensure that all students in the sample have all 3 grade level observations and each included classroom in the sample has at least 5 sample students. Note that the number of sample students in a particular classroom could deviate from the actual class size of the classroom given that students can be excluded for missing test score information. Having adequate numbers of students per classroom is necessary to identify classroom effects separately from student effects, as is required in the student fixed effect analysis. I therefore impose this sample selection criteria iteratively: (i) first dropping classrooms if there are less than 5 students, and then (ii) dropping students if they do not have all three grade level observations. Because dropping classrooms with fewer than 5 students may remove one or more of a student's grade level observations, the first sample selection criteria may require dropping additional students because these students no longer have all three grade level observations. I therefore iterate on these two sample selection criteria until the total number of observations in the sample no longer changes and each classroom consists of at least 5 students who have complete information for each grade. For the NC data, this iterative procedure required 17 iterations through the sample selection criteria.[14]

3.2. Descriptive statistics

Table 2 provides descriptive information for the full sample of students, and teachers, and schools in grades 3–5. The sample consists of 507,291 students, 223 elementary schools, 44,034 classroom, and

[13] It is possible but unlikely that the teacher proctoring the end of grade test is not the student's classroom teacher. To minimize these cases, I exclude records where the proctor is not a regular classroom teacher.

[14] See the Appendix where I show results using different sample selection criteria. Requiring more valid students per classroom and therefore dropping more classroom observations reduces precision but generally yields qualitatively similar results.

15,551 teachers. The students in the sample are 47% male, 69% white, and 25% black. 63% of students are eligible for free or reduced price free lunch. The sample selection procedures generate a sample of students with above average reading and mathematics scores, as indicated by average mathematics and readings scores of 0.18. This indicates that the students in our sample have average test scores 0.18 standard deviations above the normalized 0 average score for the entire population of NC students. The dispersion in scores for the sample is also less as indicated by the lower than 1 standard deviation in scores for the sample.

The classrooms in the sample have an average class size of 23 students, and are 64% white and about 50% male. The average experience of the classroom teacher is 11 years. The elementary school teachers in the sample are 93% female and 87% white, much higher proportions of female and white than the student population. About 29% of the teachers have a master's degree. The teachers in the sample are slightly above average in terms of licensure examination scores, with an average score 0.07 relative to a standardized mean of 0 for the whole population of NC public school teachers. About 18% of the teachers change schools during the period of the sample. This is an under-estimate of cross-school movement over an entire career given that we observe up to a maximum of 9 career years.

4. Estimation

My estimation consists of two main steps. In the first step, I use the student level reading and mathematics scores to estimate unrestricted classroom effects for each test score outcome. In the second step, I estimate various teacher models in order to decompose the classroom effects into contributions from teachers, schools, and other classroom characteristics. The Appendix provides additional details. The main advantage of the two step approach is that I can robustly estimate classroom effects in the first step without relying on a particular specification of the second step teacher quality model. This is particularly convenient here since I explore many different specifications of the second step teacher quality model.[15]

Table 3 provides descriptive statistics for our 5th grade classroom analysis sample. There are 11,333 classrooms in 156 schools with 3150 teachers. The average class size in this sample is just slightly larger than in the original sample of 3rd, 4th, and 5th grade classrooms. As expected, since I exclude teachers who leave after the first year, the average number of years of experience across classrooms is higher at 13.6 years compared to 11.5 in the original sample. The teachers in this sample are more likely to be male, but have quite similar racial distribution as the baseline. The teachers in this sample are slightly more likely to have a master's degree (31.6% vs. 29%) and have higher licensing exam scores (0.11 vs. 0.073).

Given the manageable number of school and calendar periods, I include dummy variables for schools and calendar periods, and use a standard within teacher transformation to estimate the Step 2 teacher FE model. To estimate the standard errors for the Step 2 parameter estimates I use a bootstrap procedure to take account of the multiple estimation steps. A naive bootstrap procedure of simply sampling from the administrative data sample with replacement would fail to take account of the clustered/longitudinal nature of the data. I block bootstrap sample teacher-career-classroom observations by randomly re-sampling with replacement each teacher and keeping all of the teacher's career observations and assigned students.

[15] See Ballou (2009), Ishii and Rivkin (2009), Koedel and Betts (2010), and McCaffrey et al. (2009) for recent discussions of estimation issues related to teacher quality models.

Table 2
Descriptive statistics for full sample (Grades 3–5).

Number of students		507,291
Number of schools		223
Number of classrooms		44,034
Number of teachers		15,551
	Mean	Std.
Student characteristics		
Fraction male	0.47	0.50
Fraction white	0.69	0.46
Fraction black	0.25	0.43
Fraction Hispanic	0.023	0.15
Fraction free lunch	0.63	0.48
Same race as teacher	0.70	0.46
Same gender as teacher	0.52	0.50
Math score	0.18	0.93
Read score	0.18	0.92
Classroom characteristics		
Class size	23.29	3.15
Percent white	0.64	0.26
Percent male	0.51	0.08
Experience of teacher	11.52	3.90
Teacher characteristics		
Fraction male	0.07	0.26
Fraction white	0.87	0.34
Fraction black	0.12	0.33
Fraction Hispanic	0.0019	0.044
Fraction with master's degree	0.29	0.45
Licensing exam score	0.073	0.85
Change school during sample	0.18	0.38

5. Return to teaching experience

5.1. Previous models

I first report results using previous specifications. For each of the previous specifications, I re-estimate these models on the NC 5th grade teacher sample using the mathematics outcome. Results for reading are presented below. Note that because the previous papers estimate their models on either different data samples from different school systems and/or and use different sets of conditioning variables,

Table 3
Descriptive statistics for 5th grade classroom analysis samples.

Number of classrooms	11,460
Number of schools	156
Number of teachers	3182
Classroom characteristics	
Mean class size	23.6
	(2.99)
Mean frac. white	0.66
	(0.25)
Mean frac male	0.51
	(0.08)
Mean experience of teacher	13.89
	(9.87)
Teacher characteristics	
Fraction male	0.11
	(0.31)
Fraction white	0.87
	(0.34)
Fraction with master's degree	0.32
	(0.46)
Mean licensing exam score	0.11
	(0.82)
Frac. change school during sample	0.23
	(0.42)

Notes: standard deviation in parentheses.
Source: North Carolina matched student–teacher data.

Table 4
Returns to teaching experience in mathematics (FE using Kane et al., 2008 Model).

	(1)	(2)
Student model:	Lagged Score	Student FE
Teacher model:	FE	FE
Exper. 1 year	.0708	.0812
	(.00414)	(.00349)
Exper. 2 years	.0876	.106
	(.00574)	(.00467)
Exper. 3 years	.114	.124
	(.00609)	(.0047)
Exper. 4+years	.106	.132
	(.00679)	(.0061)

Notes: This table reports estimates using the experience specification from Kane et al. (2008) applied to the NC 5th grade teacher sample (mathematics outcome). All models include school and year fixed effects and variables for classroom characteristics: class size, percentage students that are white, percentage of students that are male. In addition, the lagged score student model includes variables for student gender, white, black, or Hispanic race, and free lunch status. Standard errors are computed using a cluster bootstrap procedure described in the text. The omitted category is teachers with 0 years of experience (new teachers in their first year of teaching).
Source: North Carolina matched student–teacher data.

I cannot perfectly replicate the previous approaches nor directly test the importance of the modeling assumptions for the particular data samples they use. However, the advantage of this approach is that since I am using the same data sample for each of the model estimates, I can isolate the importance of the teacher quality modeling assumptions.

Table 4 reports results using the Kane et al. (2008) experience model (4), which includes dummy variables for the first 3 years of teaching experience and an additional variable for 4 or more years of experience. I report estimates of this model using the two different student modeling assumptions. The first column reports estimates of the model using the lagged score student model and the second column reports estimates of the model for the student FE model (See Appendix for details of these models). For each student model, I estimate a teacher FE specification. The magnitude of the estimates is interpretable in terms of standard deviations of the underlying test score measure because the test scores are normalized to have standard deviation 1. In Table 4, an estimate of 0.0708 for the "Exper. 1 Year" variable indicates that the average end of grade score for a student assigned to a teacher with 1 year of experience would be 0.0708 standard deviations higher than if assigned to a new teacher with 0 years of experience.

Using the Kane et al. (2008) experience specification, the estimates of the return to experience are slightly higher using the student fixed effects model rather than the lagged score model. The magnitude of the experience estimates is higher than found in Kane et al. (2008) results, although the estimates are not strictly comparable since they include some additional variables and estimate the model on data from New York City 4th-8th grade teachers. The qualitative pattern is the same. While there is a relatively high return to the initial year of teaching experience, the estimates appear to suggest that there are relatively small gains in teacher quality for later career experience.

Table 5 estimates another restricted experience specification using the specification from Clotfelter et al. (2007) (5) applied to the same NC data sample. These estimates are similar in magnitude to those reported by Clotfelter et al. (2007) who estimate their model on a pooled sample of NC data for teachers from grades 3-5. Like the Kane et al. (2008) estimates, these estimates suggest a high return to early experience but a relatively flat profile of later experience returns.

Fig. 1 graphs the experience returns from five different previous studies: Kane et al. (2008), Clotfelter et al. (2007), Hanushek et al. (2005b), Rockoff (2004), and Harris and Sass (2011). For each specification I use teacher and student fixed effects. Results using student lagged score models are similar. As discussed above, each of these previous specifications restricts the marginal return to experience in various dimensions, with the most severe restriction on later career experience. From the Figure it is clear that these specifications indicate a high return to early experience but a near flat return to experience in later years. The specifications provide quite similar estimates for later career experience, with the Rockoff (2004) and Harris and Sass (2011) specifications providing somewhat higher estimates, and the Clotfelter et al. (2007) and Kane et al. (2008) providing somewhat lower estimates.

5.2. Smooth parametric models

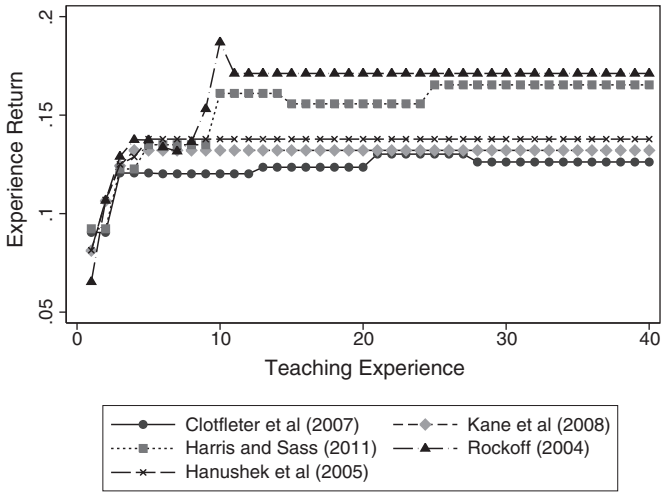
I next turn to estimating alternative experience models to test whether the experience returns estimated using previous restricted models are robust to alternative specifications. I first estimate a series of parsimonious parametric specifications, and turn to a nonparametric specification below. As with all of the previous models, I estimate these models using teacher FE and use only the within teacher variation in experience to identify the teacher quality dynamics. In the first column of Table 6, I report estimates from a linear in experience model: $f(x_{jt},\beta)=\beta_1 x_{jt}$. Unlike the previous specifications, this model allows for a smooth pattern of experience returns and does not impose that any marginal return to experience is zero. However, the linear model restricts the marginal return to experience to be a constant at $f(x)-f(x-1)=\beta_1$. I estimate the marginal return to experience in this model at 0.0362 using the student lagged score model and 0.0388 using the student FE model. Both of these estimates are statistically significant at the 1% level. The estimates imply that a teacher with 30 years of experience would have over 1 standard deviation higher level of teacher quality than a new teacher.

The remaining columns in Table 6 generalize the linear experience functions. Columns 2 and 5 add a squared experience term: $f(x_{jt},\beta)=\beta_1 x_{jt}+\beta_2 x_{jt}^2$. These estimates reveal that the return to experience is positive but concave, with more rapid gains during the

Table 5
Returns to teaching experience in mathematics (FE using Clotfelter et al., 2007 Model).

	(1)	(2)
Student model:	Lagged score	Student FE
Teacher model:	FE	FE
Exper. 1–2 years	.0782	.0904
	(.00411)	(.0033)
Exper. 3–5 years	.108	.121
	(.0057)	(.00461)
Exper. 6–12 years	.107	.12
	(.00702)	(.00582)
Exper. 13–20 years	.115	.124
	(.0087)	(.00804)
Exper. 21–27 years	.132	.13
	(.0117)	(.0101)
Exper. 28+years	.128	.126
	(.0133)	(.0113)

Notes: This table reports estimates using the experience specification from Clotfelter et al. (2007) applied to the NC 5th grade teacher sample (mathematics outcome). All models include school and year fixed effects and variables for classroom characteristics: class size, percentage students that are white, percentage of students that are male. In addition, the lagged score student model includes variables for student gender, white, black, or Hispanic race, and free lunch status. Standard errors are computed using a cluster bootstrap procedure described in the text. The omitted category is teachers with 0 years of experience (new teachers in their first year of teaching).
Source: North Carolina matched student–teacher data.



Notes: This figure plots the estimated predicted returns to experience (relative to first year teachers with 0 years of experience) using various previous experience specifications. Each estimate is from the specification listed applied to the NC data for 5th grade teachers (mathematics outcome). All estimates use teacher and student fixed effects. Experience return estimates from the student lagged score model are similar. The estimated coefficients for the Kane et al (2008) and Clotfelter et al (2007) specifications are reported in Tables 4 and 5. All models include school and year fixed effects and variables for classroom characteristics: class size, percentage students that are white, percentage of students that are male.

Source: North Carolina matched student-teacher data.

Fig. 1. Return to experience estimates using restricted FE models. Notes: This figure plots the estimated predicted returns to experience (relative to first year teachers with 0 years of experience) using various previous experience specifications. Each estimate is from the specification listed applied to the NC data for 5th grade teachers (mathematics outcome). All estimates use teacher and student fixed effects. Experience return estimates from the student lagged score model are similar. The estimated coefficients for the Kane et al. (2008) and Clotfelter et al. (2007) specifications are reported in Tables 4 and 5. All models include school and year fixed effects and variables for classroom characteristics: class size, percentage students that are white, percentage of students that are male. Source: North Carolina matched student–teacher data.

early career and slower gains in later years. As is more readily seen in Fig. 3, the concavity implied by these estimates is relatively small.

Columns 3 and 6 in Table 6 generalize the quadratic in experience model by allowing for different marginal returns to experience in the first two years of the teacher's career than in later years:

$$f\left(x_{jt},\beta\right)=\beta_1 x_{jt}+\beta_2 x_{jt}^2+\beta_3 1\left\{x_{jt}=1\right\}+\beta_4 1\left\{x_{jt}=2\right\}. \quad (8)$$

This function allows the first two years of experience to have a higher marginal return than implied by the quadratic function. For both the lagged score student model and the student FE model, I estimate that the first two years of teaching experience have higher returns than later career experience, where the marginal return to the first year of teaching is about 0.07 standard deviations. The key finding is that these estimate stands in stark contrast to the restricted experience specifications estimate above which indicate only a small or no return to later career experience.

5.3. Non-parametric in experience models

Next, I test the robustness of the previous findings from restricted parametric experience models by estimating unrestricted, non-parametric in experience models. Given that the number of years of experience is discrete with finite support, a non-parametric specification of $f(x_{jt},\beta)$ is simply a fully saturated dummy variable model with dummy variables at each level of experience:

$$f\left(x_{jt}\right)=\sum_{\tau=1}^{T}\beta_\tau 1\left\{x_{jt}=\tau\right\}, \quad (9)$$

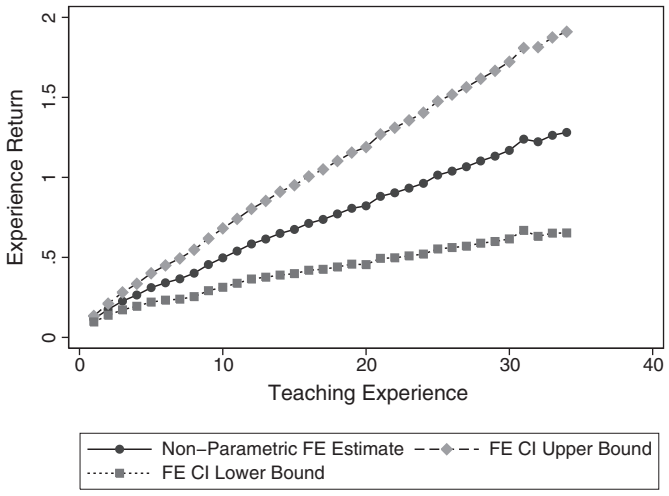
where teaching experience ranges from $x_{jt}=\{0,1,\ldots,T\}$. Fig. 2 displays the estimates and the upper and lower bounds of the 95% confidence intervals. The non-parametric pattern displays an increasing and slightly concave pattern of experience returns in teaching mathematics. The precision of the estimates is much higher for lower levels of experience in part due to the larger sample of teachers with few years of experience. Even focusing on the lower bound of the 95% confidence interval, I cannot reject at the 95% level that there is considerable later career improvement in mathematics teaching. At the lower bound, the effectiveness in teaching mathematics improves from 0.225 to 0.645 standard deviations above the mean from the 5th career year to the 30th career year, an improvement of 0.42 standard deviations or about 0.014 standard deviations per year.

Fig. 3 compares the non-parametric estimates with some of the restricted estimates. While the linear and quadratic in experience models approximate the non-parametric pattern quite well, the restricted specification (using Kane et al., 2008) does not. This demonstrates that a relatively parsimonious experience specification, using just a linear function of experience, provides a good approximation

Table 6
Returns to teaching experience in mathematics (FE models using smooth experience functions).

	(1)	(2)	(3)	(4)	(5)	(6)
Student model:	Lagged score			Student FE		
Teacher model:	FE	FE	FE	FE	FE	FE
Exper.	.0362	.0472	.0472	.0388	.0509	.0508
	(.0101)	(.01)	(.01)	(.00931)	(.00925)	(.00925)
Exper. squared		−.000327	−.000352		−.000364	−.000393
		(.0000221)	(.0000223)		(.0000209)	(.0000212)
Exper. 1 year			.0263			.0295
			(.00345)			(.00313)
Exper. 2 years			.0198			.0276
			(.00365)			(.00297)

Notes: All models include school and year fixed effects and variables for classroom characteristics: class size, percentage students that are white, percentage of students that are male. In addition, the lagged score student model includes variables for student gender, white, black, or Hispanic race, and free lunch status. Standard errors are computed using a cluster bootstrap procedure described in the text. The omitted category is teachers with 0 years of experience (new teachers in their first year of teaching).
Source: North Carolina matched student–teacher data.



Notes: This Figure reports the non-parametric estimate of teaching experience relative to new teachers (0 years of teaching experience) using teacher FE. All estimates use student FE. Experience return estimates from the student lagged score model are similar. All models include school and year fixed effects and variables for classroom characteristics: class size, percentage students that are white, percentage of students that are male. Confidence intervals are calculated from standard errors computed using a cluster bootstrap procedure described in the text.

Source: North Carolina matched student-teacher data.

Fig. 2. Non-parametric teacher FE estimate of returns to teaching experience. Notes: This figure reports the non-parametric estimate of teaching experience relative to new teachers (0 years of teaching experience) using teacher FE. All estimates use student FE. Experience return estimates from the student lagged score model are similar. All models include school and year fixed effects and variables for classroom characteristics: class size, percentage students that are white, percentage of students that are male. Confidence intervals are calculated from standard errors computed using a cluster bootstrap procedure described in the text. Source: North Carolina matched student–teacher data.

of the non-parametric relationship.[16] The key conclusion is that these restricted specifications, all using teacher FE, are very poor approximations of the teacher quality dynamics revealed by the non-parametric model.

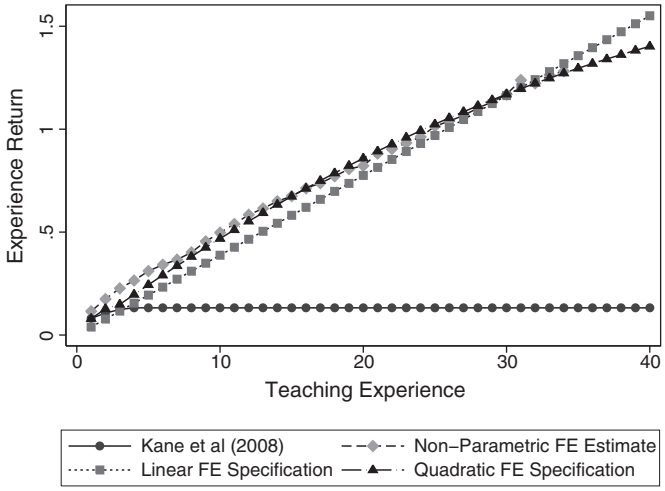
5.4. Teacher characteristics (OLS) and teacher RE models

I next examine whether the teacher characteristics model (estimated using OLS) or the teacher RE model can capture the pattern of experience returns revealed by the teacher FE specifications. The ability of these models to produce consistent estimates of the return to teaching depends on the particular distributional assumptions in the RE case and the teacher characteristics observed in the particular data at hand.

Table 7 reports estimates for the teacher characteristics (OLS) and RE models using the smooth parametric model, which was previously estimated using teacher FE in Table 6. Under the teacher RE specification, I estimate a very small return to teaching experience.[17] Using the RE estimates, a teacher with 30 years of experience would have between 0.0339 and 0.0366 standard deviations higher teacher quality than a new teacher, depending on the student model. In contrast, the return to 30 years of experience estimated using the teacher FE specification are nearly 30 times these estimates. This difference between the FE and RE model suggests that there may be substantial non-random exits from teaching (teacher self-selection) which bias downward the return to experience in the RE model relative to the more robust FE model. I return to the pattern of selection implied by the different teacher quality models below.

The teacher characteristics model I estimate includes a set of time invariant teacher characteristics, including variables for the teacher's race and gender, the teacher's initial licensing examination score (a test of general and pedagogy knowledge),whether the teacher has obtained a master's degree, the licensing status of the teacher, and whether the teacher has obtained National Board Certification. This model is estimated using OLS. Our estimates using the teacher characteristics model are quite similar to those using the teacher RE model. Like the teacher RE model, the teacher characteristics model suggests a very low return to teaching experience, especially for later career experience. This is not surprising since the previous research on teacher quality has found that many of the observable characteristics of teachers, such as education and licensure status, are only weakly correlated with teacher quality. Our findings are consistent with this conclusion.

Next, I re-estimate the teacher RE and teacher characteristics models using the restricted experience specifications as used in the previous literature. Table 8 reports estimates for teacher RE characteristics models using the Kane et al. (2008) restriction. Comparing these results with the teacher FE specification also using the same experience specification (Table 4), I see somewhat lower estimates using the RE and characteristics models relative to the FE models. For example, the first year of experience has a marginal return of 0.0708 using the teacher FE and student lagged score specification but a 0.0644 marginal return using the teacher RE specification and 0.0557 marginal return using the teacher characteristics model, also using the student lagged score model. With the restricted experience specification, the teacher FE, teacher RE, and teacher characteristics models all provide qualitatively



Notes: This Figure compares the experience return estimates (relative to new teachers with 0 years of experience) using alternative models: i) using the non-parametric estimate from Figure 2, ii) using a linear in experience model (column 4 of Table 6), iii) quadratic model (column 5 of Table 6, and iv) and the Kane et al (2008) specification (column 2 of Table 4). As seen in Figure 1 the Kane et al (2008) estimates are similar to several other previous estimates. All experience estimates use teacher FE and student FE. Experience return estimates from the student lagged score model are similar. All models include school and year fixed effects and variables for classroom characteristics: class size, percentage students that are white, percentage of students that are male.

Source: North Carolina matched student-teacher data.

Fig. 3. Comparison of experience returns under different modeling restrictions. Notes: This Figure compares the experience return estimates (relative to new teachers with 0 years of experience) using alternative models: i) using the non-parametric estimate from Fig. 2, ii) using a linear in experience model (column 4 of Table 6), iii) quadratic model (column 5 of Table 6, and iv) and the Kane et al. (2008) specification (column 2 of Table 4). As seen in Fig. 1 the Kane et al. (2008) estimates are similar to several other previous estimates. All experience estimates use teacher FE and student FE. Experience return estimates from the student lagged score model are similar. All models include school and year fixed effects and variables for classroom characteristics: class size, percentage students that are white, percentage of students that are male. Source: North Carolina matched student–teacher data.

[16] In this Figure as an example, I graph only the Kane et al. (2008) specification, which imposes that marginal return to 5th or higher years of experience is zero. From Fig. 1, it is clear that the other specifications from the previous literature are similar to the Kane et al. (2008) specification.

[17] I evaluate a fairly common specification, assuming Normally distributed teacher effects for the RE estimator.

Table 7
Returns to teaching experience in mathematics (RE and teacher characteristics models using smooth experience functions).

	(1)	(2)	(3)	(4)
Student model:	Lagged score		Student FE	
Teacher model:	RE	Charact.	RE	Charact.
Exper.	.00692	.0053	.00722	.00508
	(.000279)	(.000253)	(.000229)	(.000207)
Exper. squared	−.000193	−.000133	−.0002	−.000128
	(8.67e−06)	(7.78e−06)	(7.00e−06)	(5.95e−06)
Exper. 1 year	.0177	.00963	.0194	.0141
	(.00322)	(.00306)	(.00285)	(.00274)
Exper. 2 years	.0155	.0109	.0216	.0173
	(.00325)	(.00285)	(.00263)	(.00249)

Notes: This Table shows that random effects and teacher characteristics models underestimate the return to experience (compare to Table 6 using teacher FE. The random effects (RE) model assume teacher effect is independent of included variables. Teacher characteristics model use an OLS estimator including variables for teacher gender and missing gender, teacher race (indicators for black, Hispanic, other race, and missing race; white race is the omitted category), teacher licensing exam score and indicator missing score, whether the teacher has obtained national certification, licensing status (indicators for the teacher has an emergency or temporary license, an initial or probationary license, other type of license, or missing licensing status; regular license is the omitted category), and master's degree and missing master's degree status.
All models include school and year fixed effects and variables for classroom characteristics: class size, percentage students that are white, percentage of students that are male. In addition, the lagged score student model includes variables for student gender, white, black, or Hispanic race, and free lunch status. Standard errors are computed using a cluster bootstrap procedure described in the text. The omitted category is teachers with 0 years of experience (new teachers in their first year of teaching).
Source: North Carolina matched student-teacher data.

Table 8
Returns to teaching experience in mathematics (RE and teacher characteristics models using restricted specification).

	(1)	(2)	(3)	(4)
Student model:	Lagged score		Student FE	
Teacher model:	RE	Charact.	RE	Charact.
Exper. 1 year	.0644	.0557	.0735	.0651
	(.00396)	(.00397)	(.00332)	(.00339)
Exper. 2 years	.0763	.0637	.0901	.0751
	(.00454)	(.00425)	(.00326)	(.00297)
Exper. 3 years	.107	.0994	.111	.1
	(.00457)	(.00455)	(.00316)	(.00314)
Exper. 4+years	.0889	.0835	.102	.0883
	(.00349)	(.00355)	(.00268)	(.00263)

Notes: This table reports estimates using the experience specification from Kane et al. (2008) applied to the NC 5th grade teacher sample (mathematics outcome). The random effects (RE) model assumes teacher effect is independent of included variables. Teacher characteristics model use an OLS estimator including variables for teacher gender and missing gender, teacher race (indicators for black, Hispanic, other race, and missing race; white race is the omitted category), teacher licensing exam score and indicator missing score, whether the teacher has obtained national certification, licensing status (indicators for the teacher has an emergency or temporary license, an initial or probationary license, other type of license, or missing licensing status; regular license is the omitted category), and master's degree and missing master's degree status.
All models include school and year fixed effects and variables for classroom characteristics: class size, percentage students that are white, percentage of students that are male. In addition, the lagged score student model includes variables for student gender, white, black, or Hispanic race, and free lunch status. Standard errors are computed using a cluster bootstrap procedure described in the text. The omitted category is teachers with 0 years of experience (new teachers in their first year of teaching).
Source: North Carolina matched student–teacher data.

similar conclusions: relatively large returns to early teaching experience but small returns to later experience. But these estimates are quite different from the smooth parametric and non-parametric experience specifications. Because we do not observe this substantial difference between RE and FE models with the restricted experience models discussed above, the pattern of teacher self-selection appears to be hidden by the restricted experience specifications.

5.5. Results for reading test scores

In contrast to the results using student scores in mathematics, I do not find the same returns to later teaching experience using scores on the reading examination. Table 9 presents estimation results using the three teacher quality models (characteristics, RE, and FE) and the smooth parametric in experience model (8). For all specifications, I cannot reject the hypothesis at the 5% level that there is no return to teaching experience for reading, except for the first few years of teaching where I find some small positive return. Estimates using the non-parametric in experience model (9) are similar.[18] These estimates do reveal considerable heterogeneity in the time invariant level of teacher quality in reading, but I find no precise evidence of any dynamics in the quality of teaching with respect to reading.

An important finding is that if we were to restrict the experience model, as in the previous literature, we would see considerable similarity in results between the mathematics and reading outcomes. Table 10 re-estimates the models using the Kane et al. (2008) experience restriction (4). Results using the other restricted specifications are similar. In this table, the estimated return to experience is quite similar across the teacher quality models (teacher characteristics, teacher RE, and teacher FE models). In addition, the return to experience in reading suggested by this model is smaller but qualitatively similar to the results using the same experience restriction with the mathematics outcome (Table 4). This similarity in results could lead to a conclusion that the returns to experience are consistently small across outcomes and teacher quality models. Instead, the unrestricted models reveal that, while there is a relatively high and precise return to experience in teaching mathematics, there is a small and imprecise return to experience in reading.

Understanding why teacher experience seems to have such different returns depending on subject area is an important area for future research. My results suggest a steeper "learning curve" for mathematics than reading; mathematics teaching appears to have greater scope for improvement through experience than reading. It is important to note that teaching experience here is a "black box" and simply proxies for any human capital obtained during a teacher's career, such as learning-by-doing and more formal in-service teacher training. Harris and Sass (2011) for example find that content-focused teacher professional development appears effective in increasing student achievement but only for middle and high school math teachers, and not for reading or for earlier grades. Another study found a correlation between female teachers' own "math anxiety" and their students' math achievement (Beilock et al., 2010). Teaching experience may then be particularly effective in reducing the teacher's own math anxiety as the teacher becomes more experienced and comfortable with the mathematics curriculum.

6. Distribution of teacher quality

I next turn to what the estimated models imply about the distribution of the time invariant component of teacher quality θ_j. The variance in the time invariant teacher quality component is of central importance in the teacher quality literature as it suggests how much teacher quality actually varies in practice and provides some sense of the extent to which changes in teacher hiring and recruitment could affect student outcomes. The contribution of the present research is to show that estimates of the distribution of teacher quality critically depend on the specification of the teacher quality dynamics.

Mis-specification of the teacher dynamics can bias the estimate of the dispersion in teacher quality since the innate teacher quality component is a *residual* net of the estimated return to teaching experience. With a biased estimate of teaching experience, the innate teacher

[18] Results available on request.

quality component is also biased. In this case, I find that the downward biased estimates of the return to later career experience masks the lower innate teaching quality of long career teachers (also see below on patterns of dynamic selection). Hence, correcting the return to experience estimates reveals more heterogeneity (greater dispersion) in innate teacher quality.

Table 11 summarizes the various estimates of teacher quality dispersion for different student models and different teacher dynamics/experience models. Following the existing literature, I estimate the standard deviation of θ_j using two methods. The first method directly computes the variance in the estimated teacher effect adjusted for the variance in sampling error, and the second method uses cross-classroom (career year) correlations (See Appendix for details). The top row presents the estimates of the standard deviation of the teacher quality component estimated from the most flexible, non-parametric in experience, model (9). I estimate the standard deviation in initial teacher quality ate around 0.4 standard deviations. The remaining rows present estimates of the standard deviation using various experience restrictions from the previous literature: Kane et al. (2008), Clotfelter et al. (2007), Hanushek et al. (2005b), Rockoff (2004), and Harris and Sass (2011). My estimates from these restricted specifications of between 0.15 and 0.2 are similar to those reported in the existing literature, but are about half the size of my estimates from the non-parametric and smooth parametric models. These results show how restrictions on the experience function greatly influence the estimates of the dispersion of teacher quality.[19]

7. Dynamic teacher sorting

One of the central questions in the research on teachers is whether higher quality teachers are more likely to leave teaching.[20] I show that restrictions on the teacher dynamics have an important influence on the evidence regarding this question. I restrict attention to the sub-sample of 5th grade teachers that are observed to have left the NC public database during the 1996–2005 sample period. 1692 of the 3182 total 5th grade teachers left the NC public school teaching records during the sample period.[21] For the sub-sample of teachers I observe leaving, Fig. 4 graphs the distribution of experience at the last year before the teacher left the NC data. The distribution is bimodal, with a mass of teachers leaving in the first few years of teaching and another mass of teachers leaving around their 25th–30th year of teaching, most of whom are presumably retiring.

7.1. Sorting by estimated teacher FE

I use the estimates of the teacher effect distribution to examine how the average level of the θ_j teacher quality component varies by the timing of exit from teaching. Table 12 reports estimates of a series of regressions of teacher FE on the experience level when the teacher is observed to have exited teaching. The teacher FE is estimated using the smooth parametric experience model, which was shown to provide a good approximation to the non-parametric in experience pattern of experience returns. Column 1 of Table 12 indicates that teachers who stay in teaching an additional year have on average teacher fixed effects 0.0378 standard deviations lower than those teachers who leave. Column 2 adds a squared experience when left teaching variable. The estimates reveal that the negative association between career length and teacher quality is particularly steep in the early career years. Teachers who leave teaching after the first or second year have average mathematics teacher fixed effects about 0.5 standard deviations above the normalized test score mean of 0. Teachers who remain in teaching for 30 years have average teacher quality about 1 standard deviation below the mean. These differences are statistically significant at the 5% level (Fig. 5).

The Appendix provides a series of graphs showing the contrast in the dynamic sorting patterns across various experience models. Recall that the estimates of the innate teacher component is a residual net of estimated experience components, and biased estimates of the return to experience bias the teacher quality component. Restricting the experience model to the Kane et al. (2008) model, for example, produces a flat pattern in average innate teacher quality by year of experience when the teacher left, suggesting that there is little dynamic teacher self-selection. Not unexpectedly, the RE model reveals nearly the same pattern. The results are similar using the other experience restrictions from the previous literature. The lack of any clear pattern in sorting, in both the teacher RE and teacher FE models with the restricted experience specification, would suggest then that there is little dynamic teacher sorting: the teachers who leave teaching have about the same level of mathematics teacher quality as those that stay. In contrast, the results from Table 12 show that this conclusion is an artifact of the restrictive experience specifications.

7.2. Dynamic sorting by teacher cohort

Next I examine the extent to which the dynamic sorting patterns I document are a stationary feature of the teacher labor market or instead related to transitory teacher cohort level factors. The teacher FE estimates incorporate all time invariant teacher characteristics, and therefore reflect any differences in teacher cohort quality, where I define the teacher's cohort as the year of initial entry into teaching.[22] Previous studies (Corcoran et al., 2004a, 2004b; Bacolod, 2007) have documented a secular trend of falling quality of teachers (using measures of general aptitude and wages earned after teaching), especially among female teachers. The evidence on the reasons for teacher exits is mixed. Stinebrickner (2002) and Scafidi et al. (2006) find that most teacher exits are for family and personal reasons, for example, the birth of a child. In a series of studies using data from Michigan and North Carolina teachers, Murnane and Olsen (1989, 1990) show that teachers with higher salaries remain in teaching longer. They show that teachers with higher wage offers and higher opportunity costs of teaching, as indicated by measures of ability and college subject, are more likely to leave teaching. Dolton and von der Klaauw (1995) confirm this finding using data for teachers in the United Kingdom. Wiswall (2007) shows that teachers who leave teaching within the first few years have higher freshman college admissions SAT scores and higher non-teaching wages than teachers who leave teaching after 9–11 years of teaching.

I test the importance of cohort level changes in Table 12. Columns 3 and 4 add teacher cohort fixed effects to the regressions, thereby capturing any trend in cohort quality across the teachers. We see

[19] The Appendix plots the distribution of the estimated teacher FE. As suggested by the higher standard deviation estimate, relaxing the restriction on the teacher dynamics in the non-parametric model reveals a wider distribution of teacher quality. The higher dispersion in teacher quality is not exclusively due to outliers. Instead, we see from this Figure that relaxing the experience restriction reveals a substantially higher mass of teachers in the intermediate -0.5 to 0.5 teacher quality θ_j range.

[20] My analysis does not address the movement of teachers across schools and the distribution of teachers of differing qualities across schools. My analysis is limited to gross exits from NC public school teaching altogether. For research on how the sorting of teachers across schools see Boyd et al. (2005), Hanushek et al. (2005b), and Clotfelter et al. (2006a, 2006b).

[21] These teachers may be absent from the public school database because they have left teaching entirely for another occupation, unemployment, or some other non-employment activity (retirement or raising children); or the former NC public school teacher may be teaching in a private school, at a public school outside the state of NC, or promoted to an administrative position within the NC public school system. Using the full population of NC teachers, I calculate that about 6.26% of NC teachers who teach in 3rd–5th grade also teach in grades 6th–8th at NC public schools during the 1996–2005 period.

[22] Note that for the estimates of the return to experience using the teacher FE specifications, the source of the teacher heterogeneity, whether from cohort level differences in quality or from stationary patterns of selective exit from teaching, does not affect the robustness of these results.

Table 9
Returns to teaching experience in reading (smooth parametric model).

	(1)	(2)	(3)	(4)	(5)	(6)
Student model:	Lagged score			Student FE		
Teacher model:	Charact.	RE	FE	Charact.	RE	FE
Exper.	.00526	.00561	−.0142	.00354	.00447	−.016
	(.000359)	(.000359)	(.011)	(.000323)	(.000297)	(.00951)
Exper. squared	−.000116	−.000127	−.000169	−.0000769	−.000101	−.000177
	(.0000104)	(.000011)	(.000028)	(8.89e−06)	(8.58e−06)	(.0000241)
Exper. 1 year	.0118	.0167	.0276	.00789	.0114	.0261
	(.00439)	(.00452)	(.00503)	(.00342)	(.00353)	(.00393)
Exper. 2 years	.0124	.0139	.0192	.0123	.0119	.0166
	(.00353)	(.00356)	(.00394)	(.00325)	(.00328)	(.00381)

Notes: All models include school and year fixed effects and variables for classroom characteristics: class size, percentage students that are white, percentage of students that are male. In addition, the lagged score student model includes variables for student gender, white, black, or Hispanic race, and free lunch status. Standard errors are computed using a cluster bootstrap procedure described in the text. The omitted category is teachers with 0 years of experience (new teachers in their first year of teaching).
Source: North Carolina matched student–teacher data.

that the inclusion of these cohort fixed effects generally weakens the relationship between career length and teacher quality and I cannot reject at the 5% level that there is no statistically significant relationship. These results indicate that we cannot distinguish with any precision whether the negative dynamic selection observed is a stationary feature of the teacher labor market or a transitory feature of the particular teacher cohorts in this data.

7.3. Sorting by licensing exam score

I next turn to an alternative measure of teacher quality to see if the pattern of dynamic sorting is found using other measures. I construct a measure of initial teacher quality using the teacher's combined score on their initial licensing examinations. These licensing examinations were generally taken prior to starting the first year of teaching. Since the teachers take these exams in many different years, I normalize the exam scores to have mean 0 and standard deviation 1 among the sample of all NC public school teachers who take the exam in each year. The examinations test the teacher's general and pedagogical knowledge. Note that the teacher licensing score is not used at all in the teacher FE estimator. Since the licensing score is a time invariant characteristic of the teacher, whatever skills this test measures are absorbed in the teacher FE. Therefore, I can use the licensing score as a check on the external validity of the sorting patterns documented using the teacher FE measure.[23]

The bottom panel of Table 12 repeats the analysis using the teacher's licensing score as the dependent variable.[24] Mirroring the teacher FE pattern, the trend indicates negative selection. Teachers who stay in teaching an additional year are estimated to have 0.015 s.d. lower average scores on the licensing exam. Teachers who leave teaching after 30 years of experience have scores 1/3 of a standard deviation below the mean (normalized to zero). With teacher licensing scores, I find similar evidence of negative selection as with the teacher FE analysis. In addition, as with the teacher FE analysis, conditioning on a teacher's cohort, and using only within cohort variation in teacher exits, I find imprecise and inconclusive evidence regarding the relationship between the timing of the teacher exit and measures of teacher quality.

[23] See the Appendix for an analysis of the correlation of the licensing scores and National Board Certification and the estimated teacher FE. I generally find a positive correlation between the teacher FE and licensing scores and the teacher FE and National Board Certification, with a stronger correlation across measures for the smooth parametric and non-parametric experience models than for the restricted experience models.

[24] See Fig. 6 for the average level of the teacher licensing exam score by the year the teacher left teaching for the sample of 5th grade teachers who left teaching.

8. Discussion

8.1. Robustness

The Appendix provides several checks of the robustness of the main experience returns results. In general, we find that the results are robust to the inclusion of calendar time effects, using more restrictive sample selection criteria to include higher numbers of valid students per classroom, and to generalizations of the model allowing different returns to teaching experience by career length. In general, I find similar estimates of the linear return to experience, although different estimates of the initial return to experience. Interestingly, the teachers who have the lowest attachment to teaching (stay 4 years or fewer in teaching) have higher returns to the first and second years of teaching experience. This suggests that these teachers may have had less training prior to entering teaching and thus have a much steeper learning curve than the teachers who are more attached to teaching.

8.2. Teacher quality decomposition

At any career year t, teacher quality for teacher j, q_{jt}, is the sum of a time invariant component θ_j and a contribution from the level of experience given by $f(x_{jt},\beta)$: $q_{jt}=\theta_j+f(x_{jt},\beta)$. Fig. 7 decomposes the overall average level of teacher quality by level of experience into the respective contributions from the time invariant component and the level of teacher experience. I express the average level of teacher quality at each experience level relative to the average level for new teachers (with 0 experience). The relative average level of teacher quality at experience level x, net of school, time, and other classroom characteristics is given by

$$E[q_{jt}|x]-E[q_{jt}|0]=\left(E[\theta_j|x]-E[\theta_j|0]\right)+(f(x,\beta)-f(0,\beta)).$$

I use the unrestricted, non-parametric in experience, teacher FE specification, and student FE model to estimate the two teacher quality components. Since all of the models include school and calendar period fixed effects and classroom characteristics, the estimates of the teacher quality components reflect the contribution net of school, year, and student classroom variation. I estimate these components for the full sample of all 5th grade teachers.

Fig. 7 indicates that while the average level of teacher quality remains largely unchanged across the levels of teacher experience, the contribution of the two components has decidedly different trends. The average level of the time invariant component is trending downward, reflecting the exit of higher innate quality θ_j teachers. On the other hand, the accumulated teaching experience combined with

Table 10
Teaching experience in reading (restricted specification).

	(1)	(2)	(3)	(4)	(5)	(6)
Student model:	Lagged score			Student FE		
Teacher model:	Charact.	RE	FE	Charact.	RE	FE
Exper. 1 year	.0471	.0504	.054	.055	.0579	.0603
	(.00578)	(.00571)	(.00564)	(.00423)	(.00419)	(.00439)
Exper. 2 years	.0539	.0565	.0588	.065	.066	.0673
	(.00546)	(.00512)	(.00586)	(.00396)	(.0039)	(.00485)
Exper. 3 years	.0636	.0651	.0614	.0757	.0769	.0741
	(.00563)	(.0051)	(.00633)	(.00404)	(.00381)	(.0055)
Exper. 4+years	.0782	.0773	.0674	.08	.0846	.0897
	(.00433)	(.00393)	(.00751)	(.00328)	(.00299)	(.00639)

Notes: This table reports estimates using the experience specification from Kane et al. (2008) applied to the NC 5th grade teacher sample (reading outcome). All models include school and year fixed effects and variables for classroom characteristics: class size, percentage students that are white, percentage of students that are male. In addition, the lagged score student model includes variables for student gender, white, black, or Hispanic race, and free lunch status. Standard errors are computed using a cluster bootstrap procedure described in the text. The omitted category is teachers with 0 years of experience (new teachers in their first year of teaching).
Source: North Carolina matched student–teacher data.

estimates of a substantial return to experience is causing the experience component to trend upward. The two teacher quality components effectively cancel each other out, causing the overall level of teacher quality to remain unchanged. This pattern of negative selection and positive experience returns implies that while experienced teachers are not on average better teachers (their level of overall quality q_{jt} is not much higher than that for less experienced teachers), there is still a substantial return to later career experience. It is important to separate the concept of the return to experience from the average level of quality by experience. The return to experience is defined for a given teacher, and in the FE model is identified from within teacher changes in teacher quality: I estimate that while more experienced teachers are not on average more effective in teaching mathematics, a given teacher with a certain level of initial quality θ_j is estimated to have substantial gains in overall quality over the course of her career.

How large are these gains from experience relative to the distribution of initial teacher quality? Consider a teacher who starts at the 25th percentile of the initial teacher quality distribution in mathematics (distribution of θ_j). Given the experience returns I estimate, this teacher is expected to move to the 44th percentile of the initial distribution after 5 years experience, the 62nd percentile after 10 years experience, and the 91st percentile after 20 years experience. Note that this movement is relative to the *initial* teacher qualilty distribution (i.e. the distribution for teachers with no experience). Since the gains in experience are accumulated by all teachers if they remain in teaching, the relative placement of these teachers would not in fact change if all teachers were to remain in teaching. To provide futher context for the findings, I examine this scenario next.

8.3. Magnitude of experience returns

From the low experience return estimates and substantial dispersion in initial teacher quality, some researchers have concluded that public schools should emphasize recruitment and selection of higher initial teacher quality teachers rather than increase incentives for teacher retention. Crucial to this reasoning is the estimates of low return to later teaching experience. Ballou (1996) for example concludes that attrition from teaching is of little concern under the assumption that any gains in teacher quality are reached within the first 4 years of their career, with no later career gains in experience. Kane et al. (2008) make a similar point regarding the hiring of alternatively certified teachers (e.g. Teach for America teachers). There is little cost to the higher turnover among alternatively certified teachers since the returns to experience are quite "modest."

Like much of the education production analysis, I cannot provide direct policy recommendations from estimates of the education production function itself. However, to provide some sense of the magnitude of these estimates, I consider an experiment in which I change the exit rates from teaching. An upper bound on the importance of experience can be calculated by estimating the level of teacher quality if attrition from teaching were completely eliminated. In this scenario, assuming a stationary distribution of initial cohort sizes and a 35 year career length (up to 34 years of experience), the expected fraction of teachers with any given level of experience would be $\overline{p}_x = pr(x_{jt} = x) = 1/35$ for $x = 0,1,\ldots,34$. I contrast this with the actual cross-sectional distribution of experience in the NC public schools, estimated from the data sample given by p_x. At any given initial quality level θ_j, I can then calculate the difference in teacher quality between the "no attrition" workforce and the current workforce as

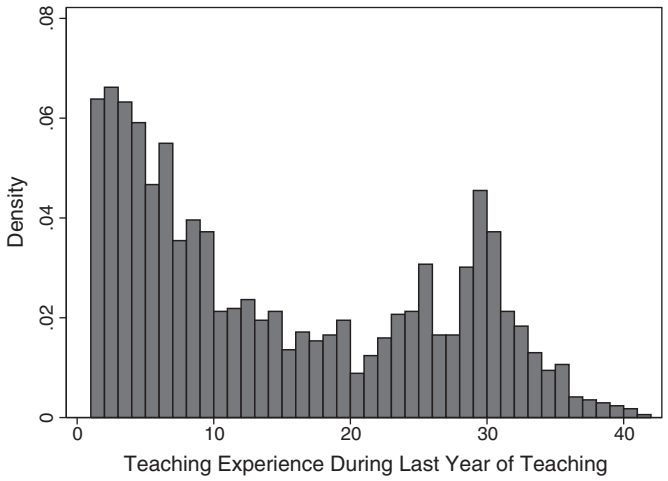
$$\overline{\Delta} = \sum_{x=0}^{34} \overline{p}_x f(x;\beta) - \sum_{x=0}^{34} p_x f(x;\beta)$$

I estimate this difference in outcomes given our non-parametric in experience estimates for $f(x;\beta)$. $\overline{\Delta}$ is estimated to be 0.34 standard deviations. An upper bound on the importance of experience to teacher

Table 11
Dispersion of teacher quality under alternative specifications.

	(1)	(2)	(3)	(4)
Dispersion estimator:	Std. teach. effect (1)		Std. teach. effect (2)	
Student model:	Lagged score	Student FE	Lagged score	Student FE
Teacher model:	FE	FE	FE	FE
Non-parametric in Exper.	.389	.398	.388	.395
Clotfelter et al. (2007) Restriction	.177	.153	.207	.19
Kane et al. (2008) Restriction	.176	.153	.207	.19
Harris and Sass (2011) Restriction	.178	.155	.208	.191
Hanushek et al. (2005b) Restriction	.176	.154	.207	.19
Linear Restriction	.408	.423	.408	.422
Quadratic Restriction	.407	.419	.405	.416
Cubic Restriction	.417	.433	.414	.429

Notes: This table calculates the standard deviation of the teacher effect under different modeling assumptions. All models include school and year fixed effects and variables for classroom characteristics: class size, percentage students that are white, percentage of students that are male. In addition, the lagged score student model includes variables for student gender, white, black, or Hispanic race, and free lunch status. Standard errors are computed using a cluster bootstrap procedure described in the text. The omitted category is teachers with 0 years of experience (new teachers in their first year of teaching). Std. Teach. Effect (1) is the standard deviation of the teacher effect calculated using the estimated variance in the teacher FE adjusted for sampling error. Std. Teach. Effect (2) is the standard deviation of the teacher effect calculated using the across classroom correlation in teacher effects. See Appendix for details.
Source: North Carolina matched student–teacher data.



Notes: This figure plots the distribution of the years of experience/career year at the last year of teaching for the sample of 5th grade teachers who left teaching during the 1997-2005 sample period. 1,692 of the 3,182 total 5th grade teachers left the NC public school teaching records during the sample period.

Source: North Carolina matched student-teacher data.

Fig. 4. Distribution of teacher experience when teacher left teaching. Notes: This figure plots the distribution of the years of experience/career year at the last year of teaching for the sample of 5th grade teachers who left teaching during the 1997–2005 sample period. 1692 of the 3182 total 5th grade teachers left the NC public school teaching records during the sample period. Source: North Carolina matched student–teacher data.

quality is that average quality would increase by about 1/3 of a standard deviation if school districts were able to provide sufficient incentives to teachers to remain in teaching for their entire working life.

Table 12
Dynamic teacher sorting.

Panel A: using teacher FE estimates				
	(1)	(2)	(3)	(4)
Dependent variable: Teacher FE				
Student model:	Student FE			
Teacher model:	FE	FE	FE	FE
Teacher cohort FE:	No	No	Yes	Yes
Experience when left	−.0378	−.0448	−.00122	.00358
	(.00925)	(.00865)	(.00102)	(.00207)
Experience when left Sq		.000212		−.000166
		(.0000326)		(.0000545)
Panel B: using teacher's licensing exam score				
	(1)	(2)	(3)	(4)
Dependent variable: teacher's licensing exam score				
Teacher cohort FE:	No	No	Yes	Yes
Experience when left	−.0154	−.0168	−.000993	−.0436
	(.00215)	(.00884)	(.015)	(.0244)
Experience when left Sq		.0000416		.00148
		(.000275)		(.000783)
Observations	1517	1517	1517	1517

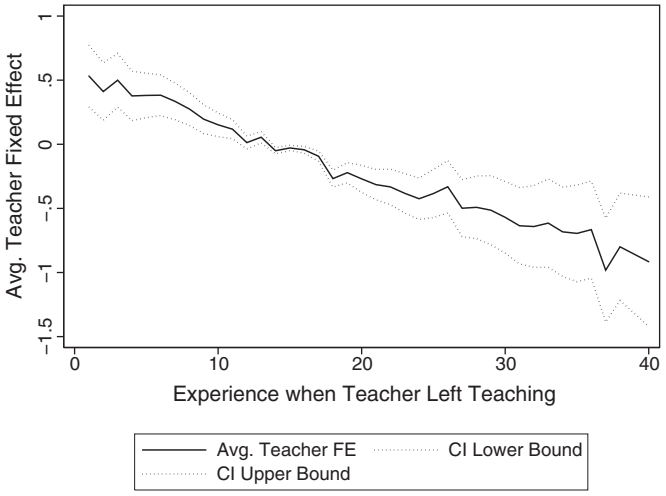
Notes: Panel A reports OLS regression estimates of the estimated teacher FE on the year the teacher left teaching for the sample of 5th grade teachers who left teaching during the 1997–2005 sample period. 1,692 of the 3,182 total 5th grade teachers left the NC public school teaching records during the sample period. The teacher effect is calculated using the teacher fixed effects specification, student FE model, and the unrestricted, non-parametric in experience specification. Panel B reports OLS regression estimates of the teacher's initial licensing exam score on the year the teacher left teaching for the sample of 5th grade teachers who left teaching during the 1997–2005 sample period. 1692 of the 3182 total 5th grade teachers left the NC public school teaching records during the sample period. The teacher effect is calculated using the teacher fixed effects specification, student FE model, and the unrestricted, non-parametric in experience specification.

This figure is similar to the standard deviation in initial teacher quality, estimated above to be close to 0.4.

While we have no direct measures of adult outcomes for this particular sample of NC 5th grade students, recent research suggests that these experience effect sizes are substantial. A 0.4 standard deviation increase in mathematics test scores from the no attrition workforce scenario amounts to about a 3% increase in scores from the mean. Chetty et al. (2011) find that a 1 percentile increase in end-of-kindergarten test socres is associated with a $94 increase in annual wage earnings at age 27 (controlling for parental characteristics). While the Chetty et al. figures are for kindergarteners in Tennessee, and my estimates are for 5th graders in NC, applying these estimates implies about a $300 increase in annual earnings under the no teacher attrition scenario. Interestingly, Chetty et al. (2011) report that the direct effect sizes for having a kindergarten teacher with 10 years experience versus having a teacher with less than 10 years experience are even larger at $1093, although, as before, these estimates are for kindergarten teachers. In addtion, because in the Chetty et al. (2011) data teachers are observed in only one career year, their estimates conflate the initial unobserved quality differences across teachers with actual experience returns, which are separated in the teacher fixed effect analysis reported here.

9. Conclusion

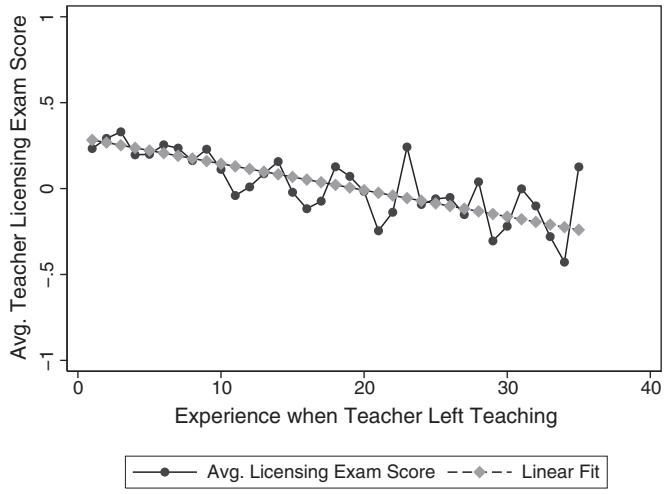
These results indicate that the basic empirical facts about the determinants of teacher quality are far from settled. Using a large sample of 5th grade public school teachers from NC, I show the importance of teacher quality modeling choices to estimates of the return to teaching



Notes: This Figure plots the average estimated teacher effect by year the teacher left teaching for the sample of 5th grade teachers who left teaching during the 1997-2005 sample period. 1,692 of the 3,182 total 5th grade teachers left the NC public school teaching records during the sample period. The teacher effect is calculated using the teacher fixed effects specification, student FE model, and the unrestricted, nonparametric in experience specification. All models include school and year fixed effects and variables for classroom characteristics: class size, percentage students that are white, percentage of students that are male.

Source: North Carolina matched student-teacher data.

Fig. 5. Teacher dynamic sorting: average teacher effect by year left teaching (unrestricted teacher FE model). Notes: This Figure plots the average estimated teacher effect by year the teacher left teaching for the sample of 5th grade teachers who left teaching during the 1997–2005 sample period. 1692 of the 3182 total 5th grade teachers left the NC public school teaching records during the sample period. The teacher effect is calculated using the teacher fixed effects specification, student FE model, and the unrestricted, non-parametric in experience specification. All models include school and year fixed effects and variables for classroom characteristics: class size, percentage students that are white, percentage of students that are male. Source: North Carolina matched student–teacher data.

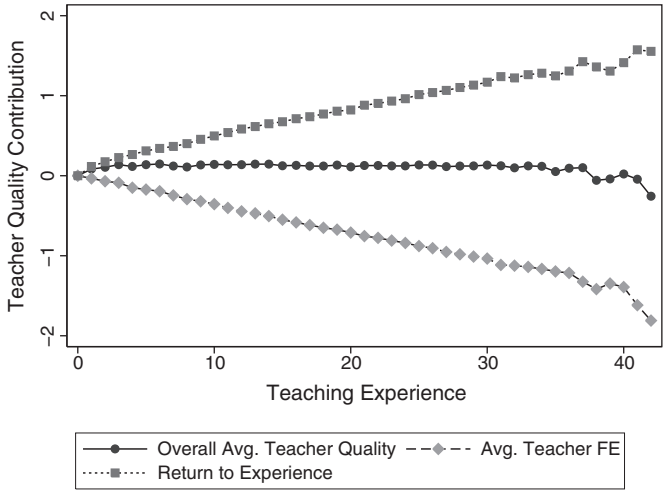


Notes: This Figure plots the average of the teacher's initial licensing exam score by year the teacher left teaching for the sample of 5th grade teachers who left teaching during the 1997-2005 sample period. 1,692 of the 3,182 total 5th grade teachers left the NC public school teaching records during the sample period.

Source: North Carolina matched student-teacher data.

Fig. 6. Teacher dynamic sorting: average teacher licensing exam score by year left teaching. Notes: This Figure plots the average of the teacher's initial licensing exam score by year the teacher left teaching for the sample of 5th grade teachers who left teaching during the 1997–2005 sample period. 1692 of the 3182 total 5th grade teachers left the NC public school teaching records during the sample period. Source: North Carolina matched student–teacher data.

experience, the dispersion of teacher quality, and the dynamics of teacher sorting. In each of these areas, less restricted models reveal substantially different patterns than what would be revealed using restricted models



Notes: This Figure decomposes the overall average level of teacher quality by level of experience into the contribution from i) the time invariant teacher quality component and ii) the experience component. I use the unrestricted, non-parametric in experience, teacher FE specification and student FE to estimate the two teacher quality components. All models include school and year fixed effects and variables for classroom characteristics: class size, percentage students that are white, percentage of students that are male. These estimated teacher quality components then reflect the contribution net of school, year, and student classroom variation.

Source: North Carolina matched student-teacher data.

Fig. 7. Decomposition of teacher quality dynamics. Notes: This Figure decomposes the overall average level of teacher quality by level of experience into the contribution from i) the time invariant teacher quality component and ii) the experience component. I use the unrestricted, non-parametric in experience, teacher FE specification and student FE to estimate the two teacher quality components. All models include school and year fixed effects and variables for classroom characteristics: class size, percentage students that are white, percentage of students that are male. These estimated teacher quality components then reflect the contribution net of school, year, and student classroom variation. Source: North Carolina matched student–teacher data.

from previous studies. For the mathematics outcome, using less restricted experience specifications and robust teacher fixed effects models, I find much higher returns to teaching experience and around twice as much dispersion in teacher quality. I do not find similar results for reading, where the lack of precision prevents a strong conclusion about the importance of later career experience to teaching effectiveness in reading. Importantly, I show that this difference in results across subject areas would be masked by restrictive models in which the estimates for reading and mathematics are qualitatively similar.

An important area for future research is to test whether these results generalize to different grade levels and other school systems. An important result is that restrictive modeling assumptions can mask substantial heterogeneity in results. I demonstrate this as I obtain similar returns to experience using teacher characteristics, teacher random effects, and teacher fixed effects models *if* I restrict the teacher experience function as in the previous literature.

A distinct issue is that estimates of the technology of education production cannot directly indicate the potential consequences of policy changes. A number of policy changes, such as increasing teacher pay, basing teacher compensation on measured performance (merit pay), and tightening licensing requirements to become a teacher, have been discussed as ways to improve the quality of teaching. However, it is unclear from the current education production function literature how these policies would affect the distribution of teacher quality. As Hanushek et al. (2005a) point out, the limitation of this education production function research is that it is "largely an input into understanding how market force play out in the teacher labor market." My results suggest a role for policy changes that would encourage greater teacher retention, through higher salaries for example. It is important to note that policies aimed at allowing for the dismissal of poor performing teachers and policies aimed at increasing retention are not mutually exclusive: the optimal teacher workforce is one comprised of high ability teachers who stay in teaching for lengthy careers.

Appendix A. Estimation with restricted experience specifications

This Appendix discusses the pooled OLS and fixed effects (FE) estimators with restricted experience specifications. Consider the following model relating teacher quality π_{jt} to innate teacher quality θ_j and teacher experience x_{jt}

$$\pi_{jt} = \theta_j + \beta d_{jt} + \phi_{jt}, \qquad (A-1)$$

where $d_{jt} = 1\{x_{jt} \geq \bar{x}\}$. As is common in the teacher quality literature, the data consists of unbalanced panel data in which for some teachers I only observe observations for a part of the teacher's career. Let T_j indicate the number of years I observe teacher j's career and Ω_j the set of teacher j's career observations. There is both truncation from below and above: for younger teachers, I observe only the first years of their career and do not observe later future years; for older teachers, the data do not include early career observations as the data collection begins after they have begun their career.

I divide the sample of teachers into i) teachers who are observed in career years *both* before and after $\bar{x}$ ("switchers") and ii) teachers who are observed with career years all before $\bar{x}$ or all after $\bar{x}$ ("stayers"). Define the within transformations: $\Delta d_{jt} = d_{jt} - \frac{1}{T_j}\sum_{t\in\Omega_j} d_{jt}$ and $\Delta\pi_{jt} = \pi_{jt} - \frac{1}{T_j}\sum_{t\in\Omega_j}\pi_{jt}$. The "switcher" teachers have $\Delta d_{jt} \neq q0$ for some t and the "stayer" teachers have $\Delta d_{jt} = 0$ for all t. Note that the delineation of "switcher" and "stayer" teacher is in reference to the particular experience specification Eq. (A-1) and is not an innate teacher characteristic; these terms do not refer to teachers who stay or switch schools.

Without loss of generality, order the J teachers in the sample so that the teachers $j=1,\ldots,J_1$ are switchers, and the teachers $j=J_1+1,\ldots,J$ are the remaining stayers. The FE estimator for β is then

$$\hat{\beta}_{FE} = \frac{\sum_{j=1}^{J}\sum_{t\in\Omega_j}\Delta\pi_{jt}\Delta d_{jt}}{\sum_{j=1}^{J}\sum_{t\in\Omega_j}\Delta d_{jt}\Delta d_{jt}} = \frac{\sum_{j=1}^{J_1}\sum_{t\in\Omega_j}\Delta\pi_{jt}\Delta d_{jt}+\sum_{j=J_1+1}^{J}\sum_{t\in\Omega_j}0}{\sum_{j=1}^{J_1}\sum_{t\in\Omega_j}\Delta d_{jt}\Delta d_{jt}+\sum_{j=J_1+1}^{J}\sum_{t\in\Omega_j}0} = \frac{\sum_{j=1}^{J_1}\sum_{t\in\Omega_j}\Delta\pi_{jt}\Delta d_{jt}}{\sum_{j=1}^{J_1}\sum_{t\in\Omega_j}\Delta d_{jt}\Delta d_{jt}}$$

This shows that the FE estimator uses only the observations from the teachers who are observed "switching" The teachers who are only observed in the data with $\bar{x}$ or more years of experience throughout their career observations or strictly less than $\bar{x}$ experience do not contribute at all to the FE estimator β.

In contrast, the pooled OLS estimator for β incorporates all teachers. The pooled OLS estimator for β is given by

$$\hat{\beta}_{OLS} = \hat{E}\left[\pi_{jt}\middle|x_{jt}\geq\bar{x}\right] - \hat{E}\left[\pi_{jt}\middle|x_{jt}<\bar{x}\right]$$

where $\hat{E}\left[\pi_{jt}\middle|x_{jt}\geq\bar{x}\right]$ is the sample mean of π_{jt} for those classrooms with a teacher who has at least $\bar{x}$ years of experience, and $\hat{E}\left[\pi_{jt}\middle|x_{jt}<\bar{x}\right]$ is the sample mean of the π_{jt} for those classrooms with a teacher who has less than $\bar{x}$ years of experience. These two sample means are computed using the entire sample of teachers. The pooled OLS estimator reflects the mean difference in classroom quality between *all* classrooms with less than $\bar{x}$ years of experience and *all* classrooms with more than $\bar{x}$ years of experience.

Appendix B. Estimation details

B.1. Main estimates

The estimation consists of two main steps. In the first step, I use the student level reading and mathematics scores to estimate unrestricted classroom effects for each test score outcome. In the second step, I estimate various teacher models in order to decompose the classroom effects into contributions from teachers, schools, and other classroom characteristics. The main advantage of the two step approach is that I can robustly estimate classroom effects in the first step without relying on a particular specification of the second step teacher quality model. This is particularly convenient here since I explore many different specifications of the second step teacher quality model. See Ballou (2009), Ishii and Rivkin (2009), Koedel and Betts (2010), and McCaffrey et al. (2009) for recent discussions of estimation issues related to teacher quality models.

Step 1 Estimating classroom effects

In the first step, I estimate classroom effects by purging the student test score outcomes of student ability and student to teacher match specific gains. The major empirical challenge is that because students are non-randomly matched to classrooms, simple OLS estimation of classroom effects are biased. I rely on the existing methodologies of using the student's prior grade test scores and student fixed effects to control for heterogeneity in student ability. I use two student models: a lagged score model, which is a generalization of a widely used value added specification, and a student fixed effect model, which assumes there is a persistent student ability component.

B.1.1. Student model 1: lagged score model

The student lagged score model specifies that the 5th grade score (either math or reading) for student i observed in classroom indexed j, t (teacher j at teacher career year t) y_{ijt} is a function of the student's specific prior (3rd and 4th grade) scores and demographic information, student to teacher match specific variables, an unrestricted classroom effect π_{jt}, and an unobserved student and classroom shock *show* ε_{ijt}.

$$y_{ijt} = Q_i'\delta + R_{ijt}'\chi + \pi_{jt} + \epsilon_{ijt} \qquad (B-1)$$

Q_i is a vector of student specific prior scores and demographic information, consisting of 3rd and 4th grade mathematics and reading scores, these scores squared, and these scores interacted (14 variables total), and demographic variables including variables for gender, race (white, black, or Hispanic), and free lunch status (a poverty indicator). R_{ijt} is a vector of student to teacher match specific variables including an indicator for whether the student is the same race as the teacher and an indicator for whether the student is the same gender as the teacher. I absorb any classroom level variation by including the π_{jt} unrestricted classroom effects (or dummy variables for each classrooms). π_{jt} reflects the contribution from schools, teachers, calendar period, and other classroom level characteristics (class size or peers). Note that I estimate separate classroom effects for mathematics and reading scores.

This lagged score model is a generalization of a commonly used teacher "value added" model. In some of these models, the left-hand side outcome is across grade differences in student outcome (e.g. grade 5 score in mathematics minus grade 4 score in mathematics). This type of model imposes a restricted depreciation pattern in student learning (Todd and Wolpin, 2003). Instead Eq. (B-1), with a flexible quadratic specification in prior scores, allows a more general relationship between prior scores and current (5th grade) scores. I use a standard FE estimator (within transformation) to estimate the π_{jt} classroom effects.

B.1.2. Student model 2: student FE

The second student level model allows for an unrestricted individual level student component to the student outcome test score. Using the panel of scores for each student (3rd, 4th, and 5th grade scores), I estimate the following student FE model:

$$y_{ijt} = \alpha_i + R_{ijt}'\chi + \pi_{jt} + \epsilon_{ijt} \qquad (B-2)$$

In contrast to lagged score student model, the student FE model assumes no relationship between lagged scores and current scores, except through the time invariant student FE component α_i. However, this model allows for an unrestricted student ability component, which may robustly capture persistent student ability if the lagged score specification is mis-specified. Because this model has both high dimensional student fixed effects and high dimensional classroom fixed effects, I cannot use standard within transformation methods or saturated dummy variable specifications to estimate the model. Instead, I use the iterative procedure developed by Arcidiacono et al. (2009), which is also used in the education production function estimation context by Harris and Sass (2011).

The Appendix provides the correlation in the estimated 5th grade classroom effects across the various models. There is a higher correlation in classroom effects across student models (lagged score vs. student FE) than across student outcomes (mathematics or reading) using the same student model. While the correlation between reading and mathematics classroom effects using the student lagged score model is 0.46, the correlation between mathematics classroom effects estimated using the student lagged model and the alternative student FE model is 0.9.

Step 2 Estimating teacher quality

The second step in the estimation takes the estimated classroom effects, estimated using either of the two student models, and estimates the particular teacher quality model. In order to emphasize the additional features of the classroom level variation, I generalize the notation so that $\pi_{j,t,s,\tau}$ is the classroom level quality for a

classroom taught by teacher j in career year t at school s during calendar year $\tau \in \{1996,\ldots,2005\}$. I estimate all teacher quality models for 5th grade classrooms and separately by outcome, and do not include additional notation for these features. Using the expanded notation, our teacher models take the form:

$$\pi_{j,t,s,\tau} = \theta_j + f\left(x_{jt},\beta\right) + H'_{jt}\psi + \eta_s + \kappa_\tau + \phi_{j,t,s,\tau} \qquad (B-3)$$

where θ_j is the innate teacher quality term, $f(x_{jt},\beta)$ is the particular function of teacher experience, which depends on the finite dimensional parameter vector β, H_{jt} is a vector of classroom characteristics, including class size, fraction white students, and fraction male students, η_s is the school specific fixed effect, and κ_τ is the calendar period fixed effect. $\phi_{j,t,s,\tau}$ is the remaining residual variation.

In order to consistently estimate the parameters in Eq. (B-3), I need to be able to separately identify the contribution of schools, calendar period, time invariant teacher quality, and time varying teacher quality from accumulated experience. Because the Step 1 estimator of classroom effects did not require identification of these contributions separately, I used the largest possible sample of students, teachers, and schools to estimate the unrestricted classroom effects. For the Step 2 estimator, identification requires that I impose additional sample selection criteria: i) I exclude schools with only one teacher per grade level during the sample period since I cannot separately identify the teacher FE from the school FE. ii) I exclude teachers with only one career observation, for which I have no within teacher variation in quality to separately estimate the teacher FE from the return to experience. iii) I only keep schools with at least 5 classroom observations over the 9 year sample period. Like the original sample of students, I iterate over these sample restrictions in order to ensure that I have at least 2 teachers per school, at least 2 teacher career observations, and at least 5 classroom observations per school (over all years). Even with these sample restrictions, it may still not be possible to separately identify the components of (B-3). In order to ensure identification, I progressively drop all classrooms for schools for which I cannot separately identify all the components from (B-3), including a non-parametric in experience specification for $f(x_{jt},\beta)$ and teacher FE.

An alternative is to drop some school, teacher, or calendar year fixed effects and estimate the model on a larger sample. This type of ad hoc model selection does not identify all of the components of Eq. (B-3) separately. The cost of this approach is that I limit our analysis to a smaller sample of classrooms. However, since I use the same sample for all models and have a large sample of data available, I can report valid comparisons across models with sufficient precision. Note that with the stronger teacher characteristics and RE assumptions, these sample restrictions to separately identify teacher, schools, calendar period, and experience are not strictly required. However, to maintain the comparability to the teacher FE model, I argue that it is important to use the same estimation sample. As an indication that the sample selection criteria is similar to previous criteria, our baseline estimates using previous models compare favorably to previously reported estimates.

B.2. Estimation methods for dispersion of teacher quality

Following the existing literature, I estimate the standard deviation of θ_j using two methods. The first method directly computes the variance in the estimated teacher effect adjusted for the variance in sampling error. For each teacher j, I form an estimate of their θ_j component using an average of their classroom quality measures, net of the other estimated components:

$$\hat{\theta}_j = \frac{1}{T_j}\sum_{t\in\Omega_j}\left(\pi_{jt} - \hat{\Phi}_{jt}\right).$$

where $\hat{\Phi}_{jt}$ are the estimated experience, classroom characteristics, and school and calendar period fixed effect components from (B-3): $\hat{\Phi}_{jt} = f\left(x_{jt},\hat{\beta}\right) + H'_{jt}\hat{\psi} + \hat{\eta}_s + \hat{\kappa}_\tau$. The variance in θ_j reflects the true variance in θ_j and variance from the residual, within classroom, error component $\phi_{j,t,s,\tau}$. Under the assumption that $\phi_{j,t,s,\tau}$ is i.i.d. within each classroom, I first estimate the within classroom error variance by averaging the within classroom variance for each teacher across all J teachers:

$$\hat{V}\left(\hat{\phi}_{j,t,s,\tau}\right) = \frac{1}{J}\sum_{j=1}^{J}\frac{1}{T_j}\sum_{t\in\Omega_j}\left[\left(\pi_{jt} - \hat{\Phi}_{jt}\right) - \hat{\theta}_j\right]^2$$

The estimate of teacher quality dispersion is then given by the difference in the overall variance in the estimated fixed effects and the estimated error variance:

$$\hat{V}_1\left(\theta_j\right) = \hat{V}\left(\hat{\theta}_j\right) - \hat{V}\left(\hat{\phi}_{j,t,s,\tau}\right)$$

The second method to estimate the variance in teacher quality uses the cross-career year (cross-classroom) correlation to estimate the variance in the teacher quality component. Under the assumption that the within classroom error is uncorrelated across classrooms, a consistent estimator for the variance in θ_j is the correlation in cross-classroom effect across all J teachers:

$$\hat{V}_2\left(\theta_j\right) = \frac{1}{J}\sum_{j=1}^{J}\hat{C}ov\left[\left(\pi_{jt} - \hat{\Phi}_{jt}\right),\left(\pi_{jt-1} - \hat{\Phi}_{jt}\right)\right]$$

Appendix C. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jpubeco.2013.01.006.

References

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

Altonji, J.G., Williams, N., 1998. The effects of labor market experience, job seniority, and job mobility on wage growth. Research in Labor Economics 17, 233–276.

Arcidiacono, P., Foster, G., Goodpaster, N., Kinsler, J., 2009. Estimating spillovers using panel data, with an application to the classroom. Working Paper.

Bacolod, M., 2007. Do alternative opportunities matter? The role of female labor markets in the decline of teacher quality, 1960–1990. The Review of Economics and Statistics 89 (4), 737–751.

Ballou, D., 1996. Do public schools hire the best applicants? Quarterly Journal of Economics 111 (1), 97–133.

Ballou, D., 2009. Test scaling and value-added measurement. Education Finance and Policy 4 (4).

Beilock, S.L., Gunderson, E.A., Ramirez, G., Levine, S.C., 2010. Female teachers' math anxiety affects girls' math achievement. Proceedings of the National Academy of Sciences 107 (5), 1860–1863.

Boyd, D., Lankford, H., Loeb, S., Wyckoff, J., 2005. Explaining the short careers of high-achieving teachers in schools with low-performing students. American Economic Review, Papers and Proceedings 95, 166–171.

Chetty, R., Friedman, J., Hilger, N., Saez, E., Schanzenbach, D., Yagan, D., 2010. How does your kindergarten classroom affect your earnings? Evidence from project STAR. NBER Working Paper.

Chetty, R., Friedman, J., Hilger, N., Saez, E., Schanzenbach, D., Yagan, D., 2011. How does your kindergarten classroom affect your earnings? Evidence from project STAR. Quarterly Journal of Economics 126 (4), 1593–1660.

Clotfelter, C., Glennie, E., Ladd, H., Vigdor, J., 2006a. Would Higher Salaries Keep Teachers in High-Poverty Schools? Evidence from a Policy Intervention in North Carolina. Sanford Institute of Public Policy, Duke University.

Clotfelter, C.T., Ladd, H.F., Vigdor, J.L., 2006b. Teacher–student matching and the assesment of teacher effectiveness. Journal of Human Resources 41 (4).

Clotfelter, C.T., Ladd, H.F., Vigdor, J.L., 2007. Teacher credentials and student achievement: longitudinal analysis with student fixed effects. Economics of Education Review 26 (6).

Corcoran, S.P., Evans, W.N., Schwab, R.M., 2004a. Women, the labor market, and the declining relative quality of teachers. Journal of Public Policy Analysis and Management 23 (3), 450–470.

Corcoran, S.P., Evans, W.N., Schwab, R.S., 2004b. Changing labor market opportunities for women and the quality of teachers, 1957–2000. American Economic Review, Papers and Proceedings 94 (2), 230–240.

Dolton, P., von der Klaauw, W., 1995. Leaving teaching in the UK: a duration analysis. The Economic Journal 105 (429), 431–444.
Flabbi, L., Ichino, A., 2001. Productivity, seniority and wages: new evidence from personnel data. Labour Economics 8, 359–387.
Hanushek, E.A., Rivkin, S.G., 2006. Teacher quality. In: Hanushek, E., Welch, F. (Eds.), Handbook of the Economics of Education, vol. 2. Elsevier.
Hanushek, E.A., Kain, J.F., O'Brien, D.M., Rivkin, S.G., 2005a. The market for teacher quality. Working paper.
Hanushek, E.A., Kain, J.F., Rivkin, S.G., 2005b. Teachers, schools, and academic achievement. Econometrica 73 (2), 417–458.
Harris, D., Sass, T., 2011. Teacher training, teacher quality, and student achievement. Journal of Public Economics 95, 798–812.
Ishii, J., Rivkin, S.G., 2009. Impediments to the estimation of teacher value added. Education Finance and Policy 4 (4).
Kane, T.J., Rockoff, J., Staiger, D.O., 2008. What does certification tell us about teacher effectiveness? Evidence from New York City. Economics of Education Review 27, 615–631.
Koedel, C., Betts, J., 2010. Value added to what? How a ceiling in the testing instrument influences value-added estimation. Education Finance and Policy 5 (1).
Lockwood, J., McCaffrey, D.F., 2009. Exploring student–teacher interactions in longitudinal achievement data. Education Finance and Policy 4 (4).
Loken, K.V., Mogstad, M., Wiswall, M., 2012. What linear estimators miss: re-examining the effects of family income on child outcomes. American Economic Journal: Applied Economics 4 (2), 1–35.
McCaffrey, D.F., Sass, T.R., Lockwood, J., Mihaly, K., 2009. The intertemporal variability of teacher effect estimates. Education Finance and Policy 4 (4).
Murnane, R.J., Olsen, R.J., 1989. The effects of salaries and opportunity costs on duration in teaching: evidence from Michigan. The Review of Economics and Statistics 71 (2), 347–352.
Murnane, R.J., Olsen, R.J., 1990. The effects of salaries and opportunity costs on length of stay in teaching: evidence from North Carolina. Journal of Human Resources 25 (1), 106–124.
Nye, B., Konstantopoulos, S., Hedges, L.V., 2004. How large are teacher effects? Educational Evaluation and Policy Analysis 26 (3), 237–257.
Olivetti, C., 2006. Changes in women's hours of market work: the role of returns to experience. Review of Economic Dynamics 9 (4), 557–587.
Ouazad, A., 2008. Assessed by a teacher like me: race, gender, and subjective evaluations. Working Paper.
Papay, J.P., Kraft, M., 2010. Do teachers continue to improve with experience? Evidence of long-term career growth in the teacher labor mark. Working paper.
Rockoff, J.E., 2004. The impact of individual teachers on student achievement: evidence from panel data. Papers and Proceedings of the One Hundred Sixteenth Annual Meeting of the American Economic Association San Diego, CA, January 3–5, 2004 (May, 2004): The American Economic Review, vol. 94, No. 2, pp. 247–252.
Rothstein, J., 2010. Teacher quality in educational production: tracking, decay, and student achievement. Quarterly Journal of Economics 125 (1), 175–214.
Scafidi, B., Sjoquist, D.L., Stinebrickner, T.R., 2006. Do teachers really leave for higher paying jobs in alternative occupations. Advances in Economic Analysis and Policy 6 (1), 1–42.
Stinebrickner, T.R., 2002. An analysis of occupational change and departure from the labor force: evidence of the reasons that teachers leave. Journal of Human Resources 37 (1), 192–216.
Todd, P.E., Wolpin, K.I., 2003. On the specification and estimation of the production function for cognitive achievement. The Economic Journal 113, F3–F33.
Topel, R., 1991. Specific capital, mobility, and wages: wages rise with job seniority. Journal of Political Economy 99 (1), 145–176.
Wayne, A.J., Youngs, P., 2003. Teacher characteristics and student achievement gains: a review. Review of Educational Research 73 (1), 89–122.
Wiswall, M., 2007. Licensing and occupational sorting in the market for teachers. Working Paper.