Productivity Returns to Experience in the Teacher Labor Market: Methodological Challenges and New Evidence on Long-Term Career Improvement

John P. Papay
Brown University
401-863-5137
john_papay@brown.edu

Matthew A. Kraft
Brown University
mkraft@brown.edu

340 Brook Street
Box 1938
Providence, RI 02912
United States of America

December 2014

ABSTRACT
We present new evidence on the relationship between teacher productivity and job experience. Econometric challenges require identifying assumptions to model the within-teacher returns to experience with teacher fixed effects. We describe the identifying assumptions used in past models and in a new approach that we propose, and we demonstrate how violations of these assumptions can lead to substantial bias. Consistent with past research, we find that teachers experience rapid productivity improvement early in their careers. However, we also find evidence of returns to experience later in the career, indicating that teachers continue to build human capital beyond these first years.

KEYWORDS
Teacher quality
Economics of education
Teacher experience

1. Introduction

Over the past decade, efforts to improve the elementary and secondary education system in the United States have focused on ensuring that all students have an effective teacher in their classroom. The debates over how to accomplish this goal have been increasingly informed by teacher effectiveness research that has blossomed in recent years with the availability of large-scale datasets that link teachers to students and test scores. These data have allowed researchers to examine central questions about the teacher labor market, including productivity dynamics– in other words, how do teachers improve their effectiveness over the course of their careers?

The extent to which teacher performance in the classroom changes with experience has both theoretical and practical implications. Better understanding this dynamic will shed light on the relationship between employee productivity and job experience, and also inform current education policy initiatives such as teacher pay, evaluation, retention, and tenure. Many analyses of the relationship between teacher experience and productivity have relied on cross-sectional data, comparing the effectiveness of teachers at different experience levels. However, this comparison does not provide a clear picture of how teachers improve over the course of their careers, largely because it ignores the issue of attrition. Even if teachers do improve with experience, we can find flat returns to experience in the cross-section if the most effective teachers leave. Thus, the extent of within-teacher returns to experience provides more relevant guidance to policymakers about teacher improvement throughout the career.

For much of the past decade, this question has been treated as settled (Rice, 2013; TNTP, 2012). Policymakers and researchers tend to believe that teachers improve rapidly during their initial years in the classroom, but that the returns to experience flatten out after the first few years of teaching. These results have become quite influential in the policy community. However, two

recent papers in this journal find otherwise, providing evidence that teachers continue to improve over the course of their careers (Harris & Sass, 2011; Wiswall, 2013).[1]

In the first half of our paper, we reconcile these divergent results by laying out explicitly the identifying assumptions that researchers have used in estimating the within-teacher returns to experience (with teacher fixed effects), given the collinearity between experience and year for nearly all teachers. We demonstrate analytically and through simulation how violations of each assumption can bias estimates, sometimes substantially. We also propose a new approach that relies on a substantively different assumption and, thus, is subject to a different source of bias. In the second half, we use data from a large urban school district to present estimates of the within-teacher returns to experience from these different models. Examining estimates from models that rely on distinct identifying assumptions provides a clearer picture of the biases in each approach and enables us to present stronger evidence about the extent of later-career returns to experience.

Like past researchers, and consistent with theory, we find that teachers in the district improve most rapidly at the beginning of their careers. However, across models, we find that teachers continue to improve, albeit at lesser rates, past their first five years in the classroom. We also find suggestive evidence of continued returns to experience throughout the career, particularly in mathematics. These results make sense, as labor economists have long observed that employee wages continue to rise with job experience. Human capital theory supports this pattern, holding that workers build skills that translate to greater productivity (Becker, 1993). Taken together, our results suggest that the question of whether teachers continue to improve with experience is at least not settled and that policymakers should temper their policies to acknowledge this reality.

[1] Given that "tenure" and "seniority" have specific meanings in the field of education, we use the term "experience" to reflect the number of years a teacher has been in the profession.

In the next section, we describe past efforts to estimate the productivity returns to teaching experience. In section 3, we describe our dataset and measures. We then articulate the key assumptions that underlie existing approaches, propose an alternative method, and discuss the bias introduced by each approach. In section 5, we present the estimated returns to teacher experience from each of these approaches in our data. We describe several threats to the validity of our inferences and our attempts to address them in Section 6. Finally, we conclude with a discussion of the economic and educational implications of this work.

2. Estimates of the Returns to Experience in Teaching

The education sector is among the few industries for which direct estimates of worker productivity are available for much of the labor force. In recent years, education economists have produced a growing body of literature that examines the productivity returns to job experience among teachers, using estimated contributions to student test score gains as a proxy for productivity (see Todd & Wolpin, 2003, McCaffrey at al., 2004, and Harris & Sass, 2006). We focus on all aspects of productivity improvement that accrue to teachers over their careers – in other words, we seek to estimate the overall effect of experience on productivity, rather than disentangling the reasons for these returns.[2] Thus, we include as "returns to experience" the effects of formal on-the-job training, informal on-the-job learning, out-of-work training (such as formal education) and any other factors that improve teacher effectiveness over time.

Most research suggests that teachers improve a great deal at the beginning of their careers (e.g., Rockoff, 2004). Fast early-career improvement in productivity is not surprising, given that

[2] There are both substantive and practical reasons for this. Substantively, we are interested in understanding how teachers improve over the course of their careers on average. Different teachers may take different paths to such improvement. Practically, many of these elements are notoriously difficult to measure. For example, in-school professional development can take many forms, only some of which are recorded. Formal education can be captured in aggregate, such as whether teachers earn a masters' degree, but we cannot distinguish finer-grained course-taking. As such, we focus on the broader question of whether teachers improve their productivity throughout their career. Finally, we find nearly identical returns to experience when we condition on teachers' formal education.

theory implies more rapid human capital development and greater investment earlier in the career (Becker, 1993). This pattern mirrors theories of the teacher career arc, where novice teachers are often characterized as simply trying to survive in the classroom as they build key classroom management skills, learn the curriculum, and add to their instructional abilities (Johnson et al., 2004). Many factors contribute to the extent of early-career productivity growth, including the availability of effective colleagues (Jackson & Bruegmann, 2009), consistency in teaching assignments (Ost, 2014), and supportive work environments (Kraft & Papay, 2014).

However, there is less agreement about the nature of returns to experience after these early years. On one hand, shirking models suggest that teachers, who face minimal oversight and enjoy strong job protections, may stop improving once they become established in their schools (Hansen, 2009). On the other, some theories of teacher career development suggest that, beyond their first few years, teachers may continue to refine their practice and gain the relationships and time to collaborate with colleagues about instruction (Huberman, 1992). Recent evidence suggests that veteran teachers can improve their instructional effectiveness if they participate in a rigorous teacher evaluation program (Taylor & Tyler, 2012), find more productive school matches (Jackson, 2013), or engage in effective on-the-job training (e.g., Matsumura et al., 2010; Neuman & Cunningham, 2009; Powell et al., 2010; Allen et al., 2011).

As Murnane and Phillips (1981) made clear, cross-sectional estimates cannot fully distinguish between true individual returns to job experience and vintage effects (i.e., average differences in quality across teacher cohorts) or selection effects (i.e., differential attrition). We focus on this question by estimating the within-teacher returns to experience using longitudinal data with teacher fixed effects. This line of work builds on Rockoff's (2004) analysis of data from two school districts in New Jersey. Rockoff finds substantial early-career returns to

teaching experience, particularly on reading test scores, but the returns to experience on all but reading comprehension scores diminish rapidly after the first few years in the classroom. More recently, Boyd and his colleagues (2008) have applied Rockoff's general approach to examine data in New York City and North Carolina, respectively, finding qualitatively similar results.

These cross-sectional and longitudinal findings have been widely interpreted as evidence that teachers do not improve their performance beyond their first few years in the classroom (Rivkin, Hanushek, & Kain, 2005). This interpretation has had a profound effect on education policy. For example, Bill Gates (2009) asserted that "once somebody has taught for three years, their teaching quality does not change thereafter." However, recent evidence suggests that teachers may improve throughout their careers. Using data from Florida, Harris and Sass (2011) find that while the largest gains in experience accrue in the first few years, there are "continuing gains beyond the first five years of a teacher's career" (p. 1). Using data on 5th grade teachers in North Carolina, Wiswall finds that "teaching experience has a substantial and statistically significant impact on mathematics achievement, even beyond the first few years of teaching" (2013, p. 62), although he finds no such returns in reading. We seek to resolve this divergent evidence by examining these approaches in more detail.

3. Dataset and Measures

3.1 Dataset

In order to examine within-teacher returns to experience, we use a comprehensive administrative dataset from a large, urban school district in the southern United States that includes student, teacher, and test records from the 2000-01 to the 2008-09 school years. This district has over 100,000 students and nearly 9,000 teachers. Student data include demographic information, teacher-student links, and annual state test results in reading and mathematics. We

standardize these test scores to interpret our estimates as standard deviation differences in student performance.[3] Because appropriate estimation of the education production function requires both baseline and outcome test data, we focus on teachers in grades four through eight. We exclude any students in atypically small classes or substantially separate special education classes.[4] Our final dataset includes more than 200,000 student-year records, representing more than 3,500 unique teachers over the 9-year panel. These students are fairly typical for an urban school district: 43% are African-American, 38% are White, and 12% are Hispanic, 10% are English language learners, and 10% are enrolled in special educational services.

Our key predictor of interest is the amount of time a teacher has spent teaching. We rely on experience as defined on the teacher salary scale. As in most U.S. public schools, teachers are paid almost exclusively based on a combination of their years of experience and their educational attainment. Although a teacher's salary experience level is a fairly reliable indicator of actual on-the-job experience, it is not perfect. We indeed see some teachers – about 5% of our sample – whose salary experience jumps more than one year in a single year.[5] As a result, we omit teachers with non-standard experience patterns from most of our models, although we investigate what happens when we include these teachers.

The teachers in this district are fairly representative of those in urban school districts

[3] Note that this standardization does not make the scales comparable from year to year because of differences in tested material and changes in the distribution of student ability from year-to-year. However, the test measure we use does not have a vertical scale that enables inferences about student growth from year-to-year.

[4] Specifically, we exclude any teacher-year in which fewer than five students had value-added estimates. We exclude any class with more than 90% of students in special education or more than 25% of students missing previous year test scores. Doing so eliminates 7% of the sample. In Appendix Table A-3a and A-3b, we explore the sensitivity of our results to these restrictions, further limiting our sample to either (a) teacher-years in which fewer than 10 students had value-added estimates or (b) teachers for whom 40 students had value-added estimates.

[5] This can result from delays in the human resources office providing appropriate credit to teachers for past teaching experience or from simple data errors. In a sensitivity analysis, we examined the consequences of this possible measurement error by focusing on teachers whom we are confident enter the district as novices. We find that the estimated within-teacher returns to experience for these teachers are in fact greater than for the overall population, suggesting that measurement error may indeed be inducing a downward bias in our results. Results are available from the authors on request.

across the country – the large majority of teachers are white women. Most have limited classroom experience, and the number of veteran teachers is relatively small. For example, only 19% of the district's teaching staff has more than 20 years of experience. In Figure 1, we present the distribution of student-year observations in our mathematics sample, showing that there are many more observations – and thus much greater precision – for teachers early in the career.[6]

4. Bias in Estimating the Returns to Experience

There are two key challenges facing researchers who seek to estimate the within-teacher returns to experience. The first involves the widely-discussed difficulties in using student achievement data to estimate teacher productivity. There are important limitations and trade-offs in specifying education production function models to estimate teacher effectiveness. We discuss these issues briefly in section 4.3 below. The second challenge involves how to specify models to estimate the within-teacher returns to experience. For teachers with standard career patterns, year and experience are collinear. This is an example of the classic age-period-cohort problem.

4.1 Returns to Experience and the Age-Period-Cohort Problem

The collinearity between year and experience within-teacher requires researchers to make identifying assumptions to separately estimate year-to-year productivity trends and returns to experience in models that include teacher fixed effects (Deaton, 1997; Rockoff, 2004). To shed light on a central piece of this challenge, we can imagine a simple data-generating process that determines the productivity of teacher *j* in year *t*:

$$(1) \qquad \pi_{jt} = \delta_j + \alpha * f(YEAR_t) + \beta * f(EXPER_{jt}) + \varepsilon_{jt}$$

Here, a teacher's effectiveness in a given year represents the sum of her initial productivity (δ_j), any productivity shocks common across teachers in a given year ($\alpha * f(YEAR_t)$), the

[6] We omit the very few teachers who ever had more than 40 years of experience. Because our sample of teachers with more than 30 years of experience is so small, we present all figures up to a maximum of 30 years.

incremental productivity teachers gain over the course of their career ($\beta * f(EXPER_{jt})$), and an idiosyncratic mean-zero error term (ε_{jt}). Note that all approaches implicitly assume that there are no interactions between experience and year – in other words, we explicitly define the year effects as average shocks common to all teachers.

We seek to fit models that will provide unbiased estimates of β. However, directly estimating a model based on equation (1) is challenging because, within teacher, experience and year are collinear, at least for teachers with standard career trajectories. Thus, all researchers seeking to estimate β must make an identifying assumption. The existing research has used three such models; we propose a fourth. Here, we lay out these four approaches, discuss their key identifying assumptions, and describe the potential bias associated with each. In short, the key distinctions across these approaches are (a) whether they make assumptions about the returns to experience profile itself and (b) what sample they use to identify key parameters.

In theory, one possibility would simply be to omit the year effects, implicitly assuming that they are random shocks by absorbing them into the error term. Rockoff (2004) recognized the serious limitations of this approach, given that many aspects of schools change over time. For example, if a district implements a policy that boosts student achievement (e.g., smaller class sizes) across all teachers in the district, within-teacher returns to experience would appear to be inflated. Rockoff (2004) developed a creative alternative. Relying on the literature, he saw the opportunity to identify year effects off of teachers with more than 10 years of experience because such teachers did not appear to become substantially more effective in cross-sectional models (Rivkin, Hanushek, & Kain, 2005). This *Censored Growth Model* explicitly assumes that there are no returns to experience after 10 years. Thus, this model requires an assumption about the functional form of the productivity-experience profile itself and restricts our inferences about

teachers' returns to experience to only the first 10 years of the career.[7]

Rockoff's (2004) innovation enables researchers to model both year effects and the returns to experience jointly, in what we call the *Censored Growth Model*:

$$(2) \qquad \pi_{jt} = \ddot{\alpha} * f(YEAR_t) + \ddot{\beta} * f(EXPER_{jt}^{CGM}) + \lambda * 1\{EXPER_{jt} > 10\} + \delta_j + \ddot{\varepsilon}_{jt}$$

Here $EXPER_{jt}^{CGM} = \{ EXPER_{jt}$ if $EXPER_{jt} \leq 10$; 10 otherwise$\}$, and we include an indicator that experience is greater than 10. We can conceptualize this model as a two-stage approach, first estimating the year effects on the sample of teachers with more than 10 years of experience and then applying these estimated year effects to a second stage equation. Because the model explicitly assumes the coefficient on the returns to experience for teachers above 10 years of experience to be zero, it essentially omits the experience effect in this first stage. This assumption produces potentially biased estimates of the year effect, as any returns to experience after year 10 will be conflated with the year effects. Thus, the mis-estimation of the year effects produces a bias in the estimated returns to experience for early-career teachers proportional to these later-career returns to experience. If the assumption holds and teachers do not continue to improve after 10 years in the classroom, this bias is zero. However, to the extent that there are any positive returns to experience after year 10, this model understates the true returns to experience. Note that, by the same logic, any negative returns to experience after year 10 would overstate the true returns to experience.

A related approach is to specify experience as a set of indicator variables that represent ranges of experience; year effects can be identified off of teachers who fall within those ranges. For example, Harris & Sass (2011) replace $f(EXPER_{jt})$ in equation (1) with dummy variables representing ranges from 1-2, 3-4, 5-9, 10-14, 15-24, and more than 25 years of experience. One

[7] In practice, one can impose different experience cutoffs (e.g. Boyd et al., 2008) but, this model must include a range over which one cannot estimate the returns to experience.

advantage of this *Indicator Variable Model* is that it enables researchers to estimate the productivity-experience profile throughout the teaching career. In practice, by using within-bin variation to estimate the year effects, the *Indicator Variable Model* relies on a similar functional form assumption. In this case, it assumes that teacher productivity does not change meaningfully within each of these experience bins.

Thus, the source of bias in the *Indicator Variable Model* is analogous to that in the *Censored Growth Model.* Year effects are estimated off of teachers in certain experience bins, but, unlike the *Censored Growth Model*, these bins occur throughout the career. Any career growth in those bins will be conflated with year effects, leading to a downward bias in the estimated returns to experience; similarly, any within-teacher declines in productivity will lead to upward bias. Here, the bias is essentially a weighted average of the within-bin returns to experience across all of the bins used in the model. The extent of bias thus depends on the nature of the bins; it is more severe if the bins include segments of the career when teachers are changing their productivity substantially. For example, if these bins include ranges early in a teacher's career, when productivity is increasing rapidly, we expect this model to introduce a substantial downward bias.

Both of these models make important contributions by estimating the within-teacher returns to experience while simultaneously accounting for year effects, but they explicitly rely on assumptions about the quantity of interest – the nature of within-teacher productivity improvement. In a recent paper, Wiswall (2013) argues that these functional form assumptions are too strong and proposes an alternative approach that uses fully flexible specifications of year and experience. For teachers with discontinuous careers, year and experience are not collinear. Such career disruptions could occur for many reasons, such as when teachers take a medical

leave, take parental leave, or leave the district for another job but then return (Stinebrickner, 2002; Scafidi, Sjoquist, & Stinebrickner, 2006). Wiswall (2013) explicitly identifies teacher experience effects off of these teachers with non-standard patterns. In what we call the *Discontinuous Career Model*, Wiswall directly fits a model akin to that in equation (1) using all teachers in the district, including those with discontinuous careers.[8]

The identifying assumption imposed by the *Discontinuous Career Model* is quite different than in the two previous models. Because teachers with standard career trajectories cannot contribute to the estimation of both year and experience effects, the available variation to estimate the within-teacher returns to experience (β) comes from teachers with discontinuous careers.[9] This is a version of the standard fixed effects assumption, where identification is based on "switchers". Here, the bias in β depends on several factors.

The first critical factor is the extent to which this group of teachers with non-standard careers represents the population of all teachers in the district, at least in their underlying true returns to experience. The subset of teachers with discontinuous careers may not represent the broader sample for many reasons – in other words, this is a question of external validity. This likely depends, in part, on the proportion of teachers with discontinuous careers. If only a small fraction of a district's teaching force falls into this category, as it does in our district, the estimated returns to experience will be based on a narrow, and possibly unrepresentative, group.

The second factor is whether the estimated returns to experience among these teachers reflect their true returns had they not experienced career disruptions. This is a question of internal validity – can the *Discontinuous Career Model* produce unbiased estimates of the

[8] Note that Wiswall (2013) uses a two-stage estimation process where he first predicts teacher-year effects and then relates those to productivity returns to experience.
[9] We can also think of this as estimating the year effects off of these teachers with non-standard career patterns, although the potential for bias remains the same.

underlying returns to experience for this subset of teachers? Here, the reason for the disruption matters substantially. There are two types of discontinuous careers: (a) teachers who take more than one year to gain a year of teaching experience because they leave the district and return, and (b) teachers who appear to have discontinuous careers because of errors in the experience variable (e.g., indicating that they gain more than one year of experience in a single calendar year). In our sample, approximately 2% of teachers have true discontinuous careers and 5% of teachers gain more than one year of "experience" in a calendar year at some point in their career.

For the first type – teachers who leave the classroom and return[10] – one important concern is that their productivity in the year in which they leave (or return) may not be representative of their overall career trajectory; for example, teachers who go on maternity or medical leave may experience negative shocks in these years. Thus, the years around which the discontinuous career happens may be particularly problematic. Any negative productivity shocks in the years surrounding the teacher's leave from (or return to) the classroom will lead to substantial bias in estimated returns to experience. Furthermore, teachers who experience the largest shocks in these years will contribute most to the estimation of the returns to experience. As a result, the estimated returns for this group may not reflect their true returns had they not experienced career disruptions.

The second type – teachers whose apparent experience increases more than one year in a single calendar year – is a larger concern, as it arises solely from data errors. For example, some teachers may have their experience level initially misclassified, leading them to gain several years of "experience" in a single year when the human resource data is corrected. These errors are particularly relevant to the *Discontinuous Career Model* because such teachers would

[10] To be clear, teachers who move to another district and then return will not have discontinuous careers if they accrue teaching experience in the other district. For these teachers, year and experience will remain collinear. In our district, teachers generally accrue salary experience if they work in another public school district in the state.

contribute substantially to the estimated returns to experience if not removed from the sample. Furthermore, although not the case in our study, if a school district denied teachers a salary step increase for poor performance, we would see teachers with the same experience level in two different years. This practice would be particularly problematic for the *Discontinuous Career Model* because experience would be endogenous for teachers with discontinuous careers.

In sum, there are two key assumptions underlying the *Discontinuous Career Model*. The first involves external validity: the group of teachers with discontinuous careers must be representative of the broader population of interest. The second involves internal validity: the career disruptions must not affect the underlying returns to experience of this group.

We propose a fourth approach that uses the full sample of teachers to estimate returns to experience without making assumptions about the functional form of these returns. As such, we require a different assumption. In a two-stage process, we use cross-teacher variation to estimate the year effects before estimating the within-teacher returns to experience. In other words, we first model productivity as a function of both experience and year effects, without teacher fixed effects. In the age-period-cohort paradigm, our first-stage approach involves estimating period effects by omitting the cohort effects. We then extract the coefficients on the year effects from the first stage ($\hat{\ddot{\alpha}}_t$) and impose them in the second stage:

$$(3) \qquad \pi_{jt} = \ddot{\alpha} * f(YEAR_t) + \gamma * f(EXPER_{jt}) + \mu_{jt}$$
$$\pi_{jt} = \hat{\ddot{\alpha}} * f(YEAR_t) + \ddot{\beta} * f(EXPER_{jt}) + \ddot{\delta}_j + \ddot{\varepsilon}_{it}$$

Here, $\hat{\ddot{\alpha}}_t$ captures any year-to-year variation in average productivity across the district other than from changes in the teacher experience distribution. Coupling these estimated year effects with teacher fixed effects allows us to estimate the returns to experience on teacher productivity ($\ddot{\beta}$) without imposing any restrictions on the functional form of experience.

This *Two-Stage Model* relies on the identifying assumption that initial teacher effectiveness (the teacher fixed effects) is not changing across years in our panel. In our first stage, the omitted variable is the teacher fixed effect. Thus, the year effects, which underpin the second stage in our analysis, will only be unbiased if, conditional on teacher experience, teacher fixed effects are uncorrelated with year: $Cov(f(YEAR_{jt}), \delta_j \mid f(EXPER_{jt})) = 0.$ If this assumption holds, the *Two-Stage Model* will recover unbiased estimates of the population returns to experience. This approach assumes that the fixed component of teacher productivity (initial ability) is uncorrelated with year, conditional on experience. For example, this assumption means that the average productivity of a novice teacher in 2000 equals the average productivity of a novice in 2009. Importantly, our assumption must only hold over the course of our nine-year panel, rather than over the thirty year window of a long-time classroom teacher's career. If the effectiveness of teachers in the district is changing over time other than through shifts in the experience distribution, our estimated year effects – and therefore our estimated returns to experience – will be biased. More rapid change will produce bias of greater magnitude.

To review, these four models rely on different identifying assumptions. The *Censored Growth Model* and the *Indicator Variable Model* require functional form assumptions about the returns to experience profile itself. The *Discontinuous Career Model* does not make any assumptions about the returns to experience profile, but instead assumes that the average returns to experience of teachers with non-standard career profiles can be estimated without bias and is representative of all teachers in the district. By contrast, the *Two-Stage Model* uses all teachers in the district to estimate the year effects. However, it assumes that there are no productivity trends in initial teacher effectiveness over time. Note that this assumption is substantively different than

that of the other approaches.[11]

4.2 Simulation

In each of these four approaches, the magnitude and direction of the bias depends on teacher labor market patterns in the district studied. In all cases, we expect the identifying assumptions to be violated, at least to some degree, in the population. The central issue is twofold: (1) to what extent are the assumptions violated, and (2) what is the magnitude and direction of the bias induced by any such violations. To illustrate these issues more directly, we complement our discussion of the potential biases with a simulation based on the data-generating process in equation (1). Using the observed patterns of teacher experience and turnover in our dataset, we generate a value of our outcome, teacher productivity, for each teacher in each year based on their experience, the year, their simulated initial effectiveness, and random error. See Appendix A for further details.

Because the bias in the *Censored Growth Model* and the *Indicator Variable Model* depends on the nature of the underlying returns to teacher experience, we create three different "true" productivity improvement profiles, displayed in Figure 2, that represent theoretically possible profiles of the returns to teacher experience. Profile A, in which productivity completely flattens after year 10, reflects the profile assumed by the *Censored Growth Model*. Profile B reflects more standard models of the productivity profile as they are monotonically positive but with diminishing marginal returns.[12] Profile C illustrates the possibility that teachers at the later stages of their careers not only cease growing but also become less effective.

Because the bias in the *Two-Stage Model* depends on trends in teacher fixed effects over

[11] For example, the other models can still produce unbiased estimates if there are trends in initial teacher effectiveness, as long as there are no later-career returns to experience.

[12] Many economic production models assume monotonic, positive growth with decreasing returns ($f'>0$, $f''<0$). Model A, with $f'=0$ over some part of the profile, is thus non-standard.

time, we create two different sets of mean-zero teacher effects. The first is uncorrelated with year, while the second induces a positive correlation between teacher effects and year.[13] Finally, because the bias in the *Discontinuous Career Model* depends in part on assumptions about the career patterns of teachers who stop out of teaching and return, we create three sets of patterns for these teachers with discontinuous careers: one in which teachers are somewhat less effective in the year they leave the district (e.g., if they have a medical problem before they go on leave), and one in which teachers are somewhat less effective in the year they return to the district (e.g., if they have an infant at home), and one with no differences in effectiveness in these years.[14]

We thus create eighteen different simulated datasets (three profiles * two sets of teacher effects * three sets of effects for teachers with discontinuous careers); in each one, we simulate an outcome for each teacher-year. We then fit each of our four models to these data. We iterate this process 1,000 times, re-creating each of the datasets and fitting the four models. We average our fitted parameter estimates to generate estimated productivity-experience profiles, and compare these estimated returns to experience to the "true" returns to experience.

4.3 Measuring Educational Productivity

We present direct estimates of the productivity returns to experience using our longitudinal student-level data. Here, a final challenge comes in measuring educational productivity itself. The assumptions underlying these models, and the challenges of using student test scores to measure teacher effectiveness, have been documented thoroughly (Todd & Wolpin, 2003; McCaffrey et al., 2004; Reardon & Raudenbush, 2009; Harris & Sass, 2006; Kane & Staiger, 2008; Koedel & Betts, 2010; Papay, 2011; Clotfelter, Ladd & Vigdor, 2006; Rothstein,

[13] We induce a correlation of approximately 0.05 to provide an illustration of the possible bias. This is on order with the observed correlations we see in our dataset, as described below.

[14] Temporary shocks before and after a teacher leaves are, on average, 0.025 standard deviations in the relevant year.

2010). Two key concerns involve the extent to which test scores capture the complex nature of a group production process and the sorting of students to teachers, both of which confound attempts to isolate teachers' contributions to student achievement (Clotfelter, Ladd & Vigdor, 2006; Rothstein, 2010). Our basic model derives from standard models in this literature. We specify an education production function that controls for baseline student characteristics and several levels of fixed effects to account for differential sorting:

$$(4) \qquad A_{it} = \alpha_g(g(A_{i,t-1})) + \beta(f(EXPER_{jt})) + \gamma X_{it} + \varphi P_{jt} + \phi S_{st} + \pi_{gt} + \tau_s + \delta_j + \varepsilon_{igjst}$$

where the outcome of interest, A_{it}, is the end-of-year test score for student *i* in year *t*. The outcome test score is modeled as a grade-specific cubic function of the student's prior year achievement, $A_{i,t-1}$, in both mathematics and reading, as well as other time-variant observable characteristics of the student (X_{it}), their peers with the same teacher (P_{jt}), and their peers in the same school (S_{st}).[15] We include school fixed effects (τ_s) to account for any time-invariant characteristics of schools, including the sorting of students and teachers to schools. Although our focus is on the assumptions underpinning the estimated teacher returns to experience, we explore the issue of student sorting and describe the sensitivity of our results to alternative specifications of this value-added model in more detail in Section VI.

In practice, given our estimation approach with student-level data, the year effects discussed above account for any difference in conditional achievement common to all students in a given year. In other words, we can think of them as accounting for any change in performance common across all teachers in a given year in the district. Given that the yearly shocks to student

[15] We include indicators for the student's gender, race, limited English proficiency and special education status, and whether the student was absent more than 10% of the year. For peer and school-level characteristics, we include the means of these predictors as well as mean prior year math and reading achievement and the proportion of students missing test scores in the past year.

achievement may vary by grade, we incorporate a full set of grade-by-year fixed effects. In most models, we specify teacher experience using a completely flexible, non-parametric specification with indicator variables for each year of experience. For the *Censored Growth Model*, we replace $EXPER_{jt}$ with $EXPER_{jt}^{CGM}$ and an indicator that teacher experience is greater than 10. For the *Indicator Variable Model*, we specify experience as a series of dummy variables representing ranges from 1-2, 3-4, 5-9, 10-14, 15-24, and more than 25 years of experience following Harris and Sass (2011). In the *Two-Stage Model*, we model student achievement as a function of these grade-by-year effects, teacher experience indicators, and all other covariates from equation (4) except teacher fixed effects in the first stage. We then constrain the grade-by-year effects (π_{gt}) to be equal to their estimated coefficient from the first stage, and estimate the model in equation (4) using these constrained coefficients.[16]

In all cases, our parameters of interest are the coefficients on the function of teacher experience (β). Importantly, our sample sizes shrink substantially for teachers with more than ten years of experience; for example, our sample includes 73 teachers at 30 years of experience. Thus, our estimates of the returns to experience later in the career are quite imprecise. However, this approach enables us to examine the returns to experience without making any assumptions about functional form. In most of our figures and tables, we illustrate the trends using linear splines in experience, with knots at 2, 5, 10, 20, and 30 years of experience. These splines fit the non-parametric results well and enable more straightforward comparisons across models, smoothing the imprecise results at higher levels of experience. The nature of our results are unchanged if we specify experience with the completely flexible dummy variables.

[16] Because we use a two-stage approach, we derive our standard errors from a clustered bootstrap approach in which we sample teachers and use all student-teacher records associated with that teacher. For other models, we present both typical and bootstrap standard errors for comparison.

5. Estimated Returns to Teacher Experience

5.1 Simulation Results

Using simulation, we can assess the predictions generated above. In Table 1, we estimate the percent bias produced by each model, across three different simulated "true" productivity-experience profiles. For the *Censored Growth Model*, the *Indicator Variable Model*, and the *Two-Stage Model*, we also present results in the case where we induce a correlation between the teacher fixed effects and year. For the *Discontinuous Career Model*, we include teachers with disruptions in their careers and induce a productivity shock before or after these disruptions. These results support our analytical assessments of possible bias described above. We present figures illustrating these trajectories in Appendix Figures B-1 and B-2.

The *Censored Growth Model* produces almost perfectly accurate estimates when the key assumption is satisfied – teachers do not improve after ten years in the classroom, as in Profile A. However, even minor violations of this assumption, where teachers experience continued returns to experience past 10 years, introduce a substantial downward bias that understates the estimated returns to experience. In Profile B, the model understates true productivity at 30 years by nearly 40%. In all profiles there is productivity improvement at some stage; as a result, the *Indicator Variable Model* substantially understates the estimated returns to experience in all of the true experience profile by as much as 68%. In both of these models, though, the degree of bias is essentially unaffected by the correlation between teacher fixed effects and year.

As expected, the extent of bias in the *Two-Stage Model* depends instead on the trend in teacher effects over time. With no trend, the *Two Stage Model* performs almost perfectly across the range of profiles. While not sensitive to differences in the underlying productivity-experience profile, it is quite sensitive to the correlation between teacher effects and year. With a positive

correlation between teacher effects and year, the model could produce a substantial downward bias – as large as 43% at 30 years with the positive correlation we impose. With a negative correlation, however, the *Two-Stage Model* would produce an upward bias.

Finally, we examine the sensitivity of our models to assumptions concerning teachers with discontinuous careers. We find that the *Discontinuous Career Model* is quite sensitive to the assumption about these teachers; the estimated within-teacher returns to experience are affected substantially by even minor shocks to the productivity of these teachers in years around their temporary separation from the district. The small negative shocks we impose result in biases that range over 50% in either direction at 30 years. Because the other three modeling approaches do not rely on this assumption, the results with productivity shocks are robust across these scenarios and we do not present them in the table or Appendix figures.

5.2 Estimated Returns to Teaching Experience from Existing Models

Our results across all four modeling approaches support one general conclusion reached in past studies: teachers improve rapidly in their first few years of teaching. In Figure 3, we display the implied experience trajectories from the *Censored Growth Model*, the *Indicator Variable Model*, and the *Two-Stage Model* in mathematics (top panel) and reading (bottom panel). Across all three models we find that teachers improve in their ability to raise student achievement by approximately 0.08 standard deviations in mathematics and 0.05 standard deviations in reading over the first five years of their career. This represents about half of a teacher's improvement in productivity in any of the models. In Figure 4, we present analogous results from the *Discontinuous Career Model*.[17] Here, to replicate Wiswall's approach, we

[17] We present these results separately in part because they rely on a somewhat different sample, including teachers with discontinuous careers. Our estimated within-teacher returns to experience are generally consistent in our other models in both the more restricted and the larger sample. In Appendix Table A-1, we present point estimates and standard errors from the coefficients in these models, in both the more restricted and larger sample. In Appendix

present the fully flexible dummy variable specification, but we find quite similar results with our splines. The implied returns to experience are substantially larger than in other models.

After the initial years, we find consistent evidence of later career improvement, particularly in mathematics, across all models. At minimum, and consistent with Harris and Sass (2011) and Wiswall (2013), this evidence suggests that the assertion that teacher productivity improvement completely stagnates after the first 3 or 5 years in the classroom is an inaccurate characterization of the average career trajectory. Instead, we find that teachers appear to improve, at least modestly, in their ability to raise student test scores well beyond their initial years in the classroom. The extent of this later career improvement is less clear, and there appear to be important differences between the returns to experience in mathematics and in reading.[18]

Overall, we find larger returns to experience in mathematics than in reading. These results match well with the value-added literature where researchers have consistently found greater variability in teachers' effectiveness in mathematics than in reading (Hanushek & Rivkin, 2010). The extent of later-career improvement is also different across the two subjects and across our models. In mathematics (Panel A), the implied profiles from three models show a relatively similar pattern, although the *Censored Growth Model* and the *Indicator Variable Model* both suggest somewhat smaller changes in productivity than the *Two-Stage Model*. The profile from *Discontinuous Career Model* does not appear to flatten out over time, suggesting that teachers continue to improve at approximately the same rate from years 29 to 30 as they did from years 2 to 3.[19] In reading, however, the estimated profiles diverge more substantially. Again, the *Discontinuous Career Model* shows rapid and sustained improvement, while the *Censored*

Figure B-3, we present a version of Figure 6 that includes the results from all four models in the less restricted sample, in both mathematics and reading.

[18] Importantly, the *Censored Growth Model* explicitly assumes no returns to experience after a given experience level - 10 years in these estimates. Thus, it cannot inform questions about career improvement after this point.

[19] These results mirror the returns to experience that Wiswall (2013) finds in mathematics.

Growth Model and the *Two-Stage Model* show more limited improvement. However, the *Indicator Variable Model* shows no improvement from year 3 to year 30.

5.3 Examining Identifying Assumptions

The validity of the inferences from these four models depends on the extent to which the identifying assumptions are met. We show that the assumptions of each model are violated in our data, some to a greater degree than others. In most cases, these violations appear to impart a downward bias on our results. For parsimony, we focus our attention on mathematics. The results in reading are comparable, and we present analogous figures in Appendix B-4.

The identifying assumption in the *Censored Growth Model* – that teachers do not improve after 10 years of experience – can be tested in at least two ways. First, we can look at the extent of improvement near the censoring point. Here, it appears that teachers continue to improve from years 5 to 10, so the assumption that they stop improving precisely at ten years is likely violated. We can conduct a more robust test by recognizing that, if teacher returns to experience are truly flat after ten years, we should arrive at similar estimates regardless of the experience range of teachers used. In Panel A of Figure 5, we present results from this model, censoring experience at 10, 15, and 20 years. The estimates derived from these different models vary substantially; shifting the cutoff from 10 to 15 or 20 years dramatically affects the implied productivity-experience profile.[20] This suggests that the returns to teaching experience are not flat after 10 years and provides additional evidence of continued career improvement.[21]

We can use similar logic to test the assumption underlying the *Indicator Variable Model*.

[20] Interestingly, our implied profile from the model with a 20-year cutoff lies below the model with a 15-year cutoff. This would result if increases in productivity beyond 20 years exceed that between 15 and 20 years.

[21] We supplement this visual analysis with a statistical test. Again, the *Censored Growth Model* assumes no differences in productivity after 10 years. As such, we modify the model by adding two dummy variables: one to indicate that teachers have between 11 and 15 years of experience and one to indicate that they have between 16 and 20 years. We reject the null hypothesis that these dummies are jointly zero in both mathematics ($F_{2,226413}$=5.95; p=0.003) and reading ($F_{2,225444}$=4.80; p=0.008). This result confirms what the figure shows – that the returns to experience are non-zero after 10 years of experience.

Here, we vary the intervals that we use to create the teacher experience bins, particularly early in the career, and present the results in Panel B of Figure 5. In the extreme case, we use a fully flexible dummy specification for the first 9 years of teaching experience.[22] Not surprisingly, as the bins get narrower, the estimated returns to experience grow steeper *and* the extent of later-career improvement increases. Again, these results suggest a violation of the key assumption on which the *Indicator Variable Model* is based and suggest that teachers do continue to improve after the first years of their career.

Next, we assess the assumption underlying our *Two-Stage Model*, that initial teacher effectiveness (δ_j), conditional on experience, is not changing over the range of our panel, or that $Cov(f(YEAR_t), \delta_j \mid f(EXPER_{jt})) = 0$. There are several reasons why this assumption may be violated. Researchers have shown that the increasing labor market opportunities for women and minorities over the past several decades and wage compression in teaching have reduced the probability that the highest-performing college graduates enter teaching (Corcoran, Evans, & Schwab, 2004; Hoxby & Leigh, 2004).[23] On the other hand, recent efforts in the district such as targeted recruitment efforts, reduced barriers to entry through alternative pathways, and improvements in teacher preparation programs may have improved the average initial effectiveness of new teachers. As a result, we cannot determine *a priori* the direction of this bias.

We examine this assumption in three ways. First, we fit models that include a set of teacher characteristics, such as indicators of a teacher's race, gender, certification pathway, and college selectivity, in addition to experience in the first stage. Here, our estimated year effects

[22] As seen in Appendix Table A-2, we can reject the null hypothesis that each dummy variable for early-career experience is zero. In addition, we compare this more flexible model to the basic specification. Here, we see in mathematics (but not in reading) a significant difference between the estimates within bins. For example, we find that productivity in the second year is statistically different from that in the first ($F_{1,266407}$=30.14; p<0.001). This suggests that the less flexible model indeed obscures some within-bin returns to experience early in the career in mathematics that would bias downward the results.

[23] Importantly, it is less clear to what extent average initial teacher effectiveness has changed.

are purged of effects from changing demographics of the teaching work force over time. Our results with these models are nearly identical to the primary results presented above, suggesting that any changes in teacher effectiveness that would affect our results must be uncorrelated with trends in these teacher demographics. This test is necessarily weak, though, because observable teacher characteristics are not strong predictors of teacher effectiveness (Rockoff et al., 2011).

Second, we examine explicitly the covariance between observable teacher fixed effects and year, conditional on experience. We begin by fitting the basic value-added model in equation (4), including a full set of experience dummies but excluding the year effects. We estimate a fixed effect, conditional on experience, for each teacher across the full panel of data. We then regress these estimated teacher fixed effects on a set of year indicators. The joint F-test on these year indicators enables us to examine whether teachers' (estimated) initial effectiveness is changing over time. In mathematics, we find no evidence that our assumption has been violated ($F_{(9,10297)} = 0.61$; $p=0.79$). However, in reading this test rejects our null hypothesis ($F_{(9,10463)} = 5.09$; $p<0.0001$). We fine a modest positive correlation ($r=0.061$, $p<0.0001$), suggesting that the initial effectiveness of English teachers is improving over time. As seen in the simulation, this correlation would be sufficient to introduce a moderate downward bias in our results in reading.

In the test above, we exclude year effects when estimating teacher fixed effects to avoid partialling out any true differences in effectiveness correlated with year, which is what we seek to examine. However, excluding year effects may bias our estimated teacher effects. Ideally, we are seeking an unbiased absolute measure of teacher's true initial effectiveness that can be compared across our panel. Although such a measure is not available, we can use information about teachers before they enter the classroom, such as the selectivity of their undergraduate institution, as a noisy proxy for effectiveness that is uncorrelated with both year and experience.

We can thus interpret any trend in college selectivity within experience level over time as evidence of bias in our model. We regress the Barron's selectivity ranking of each teacher's undergraduate institution on year, controlling flexibly for teacher experience.[24] We find small but statistically significant relationships in both mathematics (-0.0098; p=0.032) and reading (-0.0133; p=0.006). The magnitude of these relationships suggests that the competitiveness of teachers' undergraduate institutions improves by 0.1 rating point every 10 years. The direction of this trend again suggests improvements in initial effectiveness that could induce a downward bias in the estimated returns to experience from this model.

Finally, the key assumption for the *Discontinuous Career Model* involves the sample used. In Panel C of Figure 5, we compare the results from our "full" sample, which excludes teachers who have discontinuous careers because they gain more than one year of "experience" in a single calendar year, to one that includes all teachers in the dataset, even those with data entry errors in experience. Clearly, the results from these two models are quite different, suggesting that the construction of the sample of teachers with discontinuous careers matters a great deal. The identification off of data entry errors in the full sample leads to a substantially different inference about teachers' returns to experience.

Given that the returns to experience in the *Discontinuous Career Model* are identified off a small and potentially unrepresentative sample of teachers – those who leave the classroom and then return – we test the underlying assumption in more detail. In particular, we examine whether teachers with discontinuous careers experience productivity shocks in the years before or after they return, given their overall career trajectories. In other words, we fit modified versions of our main *Discontinuous Career Model*, but also include a predictor that indicates

[24] The Barron's ranking ranges from 1 ("most competitive") to 5 ("least competitive"). We observe ratings for approximately 85% of teachers in our sample.

whether the teacher left the district after the current year or returned to the district in the current year. The coefficients on these predictors indicate whether teachers with career disruptions experienced productivity shocks in these years. In mathematics, at least, we find evidence of shocks that suggest the potential for substantial bias: teachers are less effective by 0.030 standard deviations (p=0.035) in the years they leave and 0.020 standard deviations (p=0.181) in the year they return, compared to their productivity in other years.

5.4 Returns to Experience Estimates across Models

While all four models are subject to some types of bias, taken together they provide a more complete picture of the productivity returns to teaching experience. In Figure 6, we present the results from what we consider to be the most robust specifications of the three models that identify estimate using teachers with standard career trajectories: the *Censored Growth Model* censored at 20 years of experience, the *Indicator Variable Model* with dummy variables for experience early in the career, and the *Two-Stage Model*. Despite somewhat imprecise estimates after 10 years of experience, we find consistent evidence for later-career productivity improvements across nearly all models, particularly in mathematics. Figure 4 represents our preferred specification of the *Discontinuous Career Model*, although we interpret these results more cautiously given the sensitivity of the model to sample construction (e.g. Figure 5 Panel C) and the uniquely steep and linear returns to experience profile it produces.

In Table 2, we summarize the implied returns across different ranges of experience from these models. Here, three key patterns emerge. First, as noted above, we find large and statistically significant early-career returns to experience across models in both mathematics and reading. Second, we find consistent evidence of growth in later stages of the teaching career, particularly in mathematics. From year 5 to year 15, in mathematics, we find statistically

significant improvements in teacher effectiveness between 0.033 and 0.051 standard deviations. These estimates imply returns over this 10-year period of approximately 45% to 60% of the effectiveness teachers gain in their first five years. In reading, we see less consistent evidence. Both the *Censored Growth Model* and the *Two-Stage Model* show improvements of 0.022 to 0.032 standard deviations from years 5 to 15, but these estimates are not statistically significant. In both subjects, the *Discontinuous Career Model* shows substantial improvement, implying returns of 108% in mathematics and 101% in reading over the same period. Third, the point estimates from most models suggest continued returns to experience after 10 years, particularly in mathematics, but here our limited statistical power yields very imprecise estimates. Results from the *Two-State Model* and the *Indicator Variable Model* suggest gains of approximately 0.03 SD from years 10 to 25, but only the Indicator Variable Model estimates are significant.

These returns to experience are substantial, particularly relative to other correlates of teacher effectiveness. Past research has found that very few observable teacher characteristics predict future performance (Wayne & Youngs, 2003; Rockoff, et al., 2011). The predictive power of those few characteristics that are related to teacher effectiveness is very small. For example, average differences across licensure type are commonly found to be less than 0.03 standard deviation (Kane, Rockoff, & Staiger, 2008), while National Board Certification (Clotfelter, Ladd, & Vigdor, 2007) and performance on a test of mathematical content knowledge (Rockoff et al., 2011) are associated with positive differences of approximately 0.02 standard deviations. Our estimates of returns to experience that teachers accrue after five years on the job are comparable or even larger than these teacher characteristics commonly used in the teacher hiring processes.

6. Threats to Validity

6.1 Sample Attrition

Attrition from teaching – and from the district – presents another potential challenge. Many teachers choose to leave teaching and thus are censored from the dataset. Using teacher fixed effects to focus on within-teacher variation helps to resolve the challenge posed by censoring by accounting for any differences in underlying teacher characteristics related to the probability of attrition. For example, if a teacher's decision depends on her *level* of effectiveness, our estimates will not be biased because we are only examining within-teacher returns to experience; including the teacher fixed effect alleviates this concern. But, to the extent that teachers' decisions to leave are related to their improvement in effectiveness over time that is uncorrelated with fixed teacher attributes, attrition may be problematic. If so, our estimates will reflect the returns only for those teachers who stay in the district.

Although we can never know how teachers who left the district would have performed had they stayed, we can use the best information available – their returns to experience during their time in the district – to quantify the nature of this attrition challenge. If teachers who leave have been improving at different rates than teachers with the same level of experience who stay, our estimates likely do not reflect the overall returns to experience for all teachers. We first estimate teacher productivity in each year using a modified version of equation (4) where we include teacher-year effects instead of teacher effects and we omit teacher experience. From these teacher-year effects, we calculate two different measures of productivity changes described below. We also define an indicator ($ATTRIT_{jt}$) for whether teacher *j* left the district after year *t*. We then fit models of the following form in a teacher-year dataset:

$$(5) \qquad \Delta PRODUCTIVITY_{jt} = \tau ATTRIT_{jt} + \beta * f(EXPER_{jt}) + \mu_{jt}$$

If our estimate of τ is statistically significant, it suggests sample attrition may be driving the

results we find.

We calculate two different measures of recent changes in productivity to include as outcomes. First, we compare the change in estimated productivity for each teacher from year *t-1* to year *t*. This simple method provides us with the largest possible sample and is a reasonable estimate of the instantaneous change in productivity for teachers at that level of experience. However, this method ignores the possible correlation between productivity in year *t* and a teacher's decision to leave. For example, having a challenging group of students in a particular year may affect estimates of a teacher's effectiveness as well as her decision whether to return the following year. Teachers who are planning to leave at the end of the year may also "check out", reducing their effort. Or, teachers who face some sort of external shock such as a health issue, the sudden illness of a loved one, or a divorce may not perform as well and may be more likely to leave teaching. In any of these situations, our estimates of teacher productivity improvement would be biased and our results may falsely suggest that teachers who leave would have improved less quickly than those who stay. To address these possible issues, we also look at the lagged change in productivity from year *t-2* to *t-1*. Unfortunately, this measure is unavailable for teachers who leave after only two years on the job.

Importantly, most teachers who leave the district do so in the first few years. For example, 26% of first-year teachers leave the district each year, compared to 10% of teachers with 10 to 20 years of experience. We can obviously say nothing about the potential returns to teaching experience for teachers who leave the classroom after their first year. However, for other teachers, we find only limited evidence that the returns to experience differ between teachers who leave and those who stay. In Table 3, we present estimates of τ from equation (5). In the top row, we see that teachers who leave the district may be improving over the past year at

a somewhat slower rate than those who stay, particularly in mathematics. However, given our concern about a possible negative shock in a teacher's final year that is correlated with their decision to leave, we also examine the lagged measure of change in productivity. In both mathematics and reading, we find that the estimates are very nearly zero when we use productivity changes from time *t-2* to *t-1* as our outcome, suggesting that teachers who leave do not have different long-term trajectories than those who stay.

We also examine heterogeneity in these differences across levels of experience. For example, early career teachers who leave the district may be improving more slowly than their peers who stay, but mid-career teachers who leave may be those who are improving more rapidly and possibly have better outside opportunities. If so, our estimates may be biased differently for teachers with different levels of experience. We test this hypothesis by including the interaction of $ATTRIT_{jt}$ and a full set of experience dummies and conducting a set of General Linear Hypothesis tests on these interaction terms. As seen in the bottom panel of Table 1, we find no evidence that the difference in returns to experience between leavers and stayers varies by teacher experience. Thus, a very cautious reading of these results would suggest that teachers who leave the district may be improving at slightly lower rates than those who stay, indicating that our estimated returns to experience may slightly overstate the returns for all teachers, including those who leave. These differences, however, are not large enough to change our substantive conclusions.

6.2 Sorting of Students to Teachers and Specification of the Educational Production Function

Although recent research suggests that our education production function modeling approach is robust to a variety of potential threats (Chetty, Friedman, & Rockoff, 2014), we cannot be sure that we have fully accounted for every threat posed by student sorting. This

challenge is particularly important for our work because more experienced teachers tend to teach more advantaged students (Clotfelter, Ladd, & Vigdor, 2006). This pattern holds in our sample as well, especially for novices. For example, in our sample, novice teachers teach students with past test scores that are 0.20 standard deviations lower than teachers with more than ten years of experience in mathematics, and 0.19 standard deviations lower in reading. This differential sorting of students to teachers based on teacher experience is thus an important threat.

Fortunately, several factors mitigate against this challenge in our analysis. First, all of our results derive from models with teacher fixed effects. As such, cross-sectional differences in student characteristics by experience overstate the challenge we face, which is the potential of differential sorting within teachers over the course of our panel. Second, our key inferences about returns to experience come in the later stages of teachers' careers. For these later-career results to be affected, we would need the same teacher to systematically teach different types of students after year ten. Not surprisingly, we find much less evidence of differential sorting among teachers after the first few years in teaching. In fact, among teachers with more than ten years of experience, we find no relationship between past student performance and experience. If we regress past student test scores on a linear term for teacher experience for teachers with more than ten years of experience, the coefficient on experience is close to – and not statistically distinct from – zero in both mathematics (0.0007) and reading (-0.0002). Thus, the sorting of teachers to students does not appear to drive our estimates of later career improvement.

Furthermore, some of the sorting that does exist is driven by the sorting of students and teachers to schools. For example, within schools, novice teachers are assigned students whose past test scores are 0.10 standard deviations lower than more experienced teachers in mathematics and reading, a reduction of almost half the magnitude observed across schools. In

our preferred specification, we include school fixed effects that explicitly compare teachers in the same school. We also test the sensitivity of our results by including school-level averages of student characteristics to account for differences across schools.

We include a wide range of student, peer, and school characteristics in our educational production function in equation (4) in attempt to address these sorting challenges. Here, we examine the sensitivity of our results to our decisions in specifying these models. Past assessments of the importance of modeling choices find that results are rather insensitive to many such decisions, except the decision to include school fixed effects (McCaffrey et al., 2004; Harris & Sass, 2006). However, as explained by Todd and Wolpin (2003), our strategy in equation (4) may prove problematic because a student's prior year achievement is measured with error. Given that the purpose of the controls is to account for non-random sorting of teachers to students, the processes schools use to assign students to teachers determine the appropriate correction. To the extent that schools use test scores, themselves, in the assignment process, equation (4) offers an appropriate specification. If they use other proficiency measures (such as academic grades or teacher recommendations), of which test scores are a noisy measure, then the issue of measurement error in an independent variable may prove problematic.

We follow Todd & Wolpin (2003) and Jackson & Bruegmann (2009), using a twice-lagged outcome to instrument for the once-lagged outcome. Given that the sample with twice-lagged outcomes is necessarily restricted, we take the coefficients on the lagged test scores from the 2SLS model and use them as coefficients on the lagged scores in the full sample (Jackson & Bruegmann, 2009). We find nearly identical results using this approach, although we prefer our standard model because of increased sample size and precision.

7. Discussion and Conclusion

In this paper, we contribute to the literature on the productivity returns to experience in several ways. We describe the identifying assumptions of three modeling approaches used to estimate the within-teacher returns to experience, and we introduce a fourth approach. We then document how violations of the assumptions underlying these four models can reconcile the divergent evidence they produce. In our dataset, three of these approaches appear to produce downwardly-biased estimates of the within-teacher returns to experience, and in some models, this bias is quite substantial. We find consistent evidence across models that teachers improve most rapidly during their first several years on the job but also continue to improve their ability to raise student test scores beyond the first five years of their careers. This directly contradicts the standard policy conclusion that teachers do not improve after the first three to five years of their career. Finally, we find suggestive evidence across multiple modeling approaches that teachers continue to improve even later in their careers, particularly in mathematics.

Our findings have several important implications for research and policy. They illustrate how collinearity in fixed effects models requires careful attention to potential sources of bias. While "switchers" in these models often provide useful sources of variation, at times such variation is oddly constrained and these switchers may not reflect the broader inferences of interest. This has implications not only for models that seek to include experience, year, and teacher fixed effects, but also, for example, grade, year, and student fixed effects, because grade and year are collinear within students who follow traditional course trajectories. Of course, year and experience (or grade and year) are not in and of themselves collinear – they only become issues when included in models with teacher (or student) fixed effects. Researchers should be aware of these challenges given that attempts to reduce potential biases by including increasingly fine-grained sets of fixed effects can, in some cases, introduce new biases.

Our results also point to three key extensions for future research. One, despite using data from a relatively large school district, we do not have sufficient statistical power to detect even relatively substantial later-career returns to experience. Using the estimates and standard errors from our *Two-Stage Model* in mathematics, our results suggest that we would need a sample of teachers that is approximately four times larger to detect such the returns to experience we find between years 10 and 25 at traditional levels of significance. Fortunately, given increasing availability of large-scale datasets, achieving the power necessary to detect effects would be possible using panel data from the largest districts, such as New York City, Los Angeles, or Chicago, or mid-size to large states.

Second, the district we examine is only one of many large urban districts in the country. The pattern of returns to experience might well be different in other contexts with different local teacher labor market conditions. Other districts with different policies and professional development programs may not demonstrate returns that are as great, or they may in fact be greater elsewhere. Exploring these relationships in different contexts, with larger datasets, would provide fruitful guidance to policymakers.

Third, our focus on average trends also likely obscures substantial heterogeneity in teacher productivity-experience profiles. Individual teachers necessarily have distinct profiles, which result from personal characteristics and the interaction of these individuals with their colleagues and their school context. Some organizations likely provide the conditions under which employees can continue to develop, while others do not (Kraft & Papay, 2014). Understanding the characteristics of employees, colleagues, and organizations that best promote continued productivity improvement should remain a high priority for researchers.

Nonetheless, our results provide strong evidence of average productivity growth after five years and, at a minimum, indicate that the nature of later-career returns is not a settled question as has been assumed in the literature. The patterns that we find are largely consistent with results from the broader economic literature that employee wages rise with job tenure. It is also likely an understatement of the true returns to experience for several reasons. First, in our data the identifying assumptions required across multiple models appear to be violated in ways that impose a negative bias, particularly in reading. Second, our measure of productivity is necessarily limited. Schooling is a group production process where many teachers contribute to student outcomes, and raising student test scores is only one important educational outcome. For example, Carrell & West (2011) find that in higher education, more experienced professors have less success in promoting student short-term test-score growth than their less experienced colleagues, but they contribute substantially more to their students' lasting knowledge and academic skills. Finally, particularly as schools become more collaborative workplaces, peer productivity spillovers are increasingly important. For example, using data from North Carolina, Jackson & Bruegmann (2009) find that a one standard deviation difference in average peer productivity is associated with a 10 to 20% increase in a teacher's own effectiveness. Given that our models do not account for any productivity spillovers or other effects of veteran leadership in schools, they likely understate the total returns to experience.

Our findings that teachers continue to improve in their productivity beyond the early stages of their career and, at least suggestively, throughout their career are striking for several reasons. There has been substantial debate over the extent to which rising wage-experience profiles reflect improvements in employee productivity, particularly for employees after their first years on the job. In the district we study, the wage-experience profile remains relatively

linear throughout the career, while we see clearly diminishing marginal productivity returns to experience. Teacher compensation contracts are likely doing many other things than simply rewarding productivity such as encouraging loyalty. However, our findings of continued returns to experience suggest that at least part of the observed relationship between wages and experience may reflect true productivity improvement.

Education policymakers regularly argue that the research literature is conclusive on this topic: teachers do not improve in their ability to raise student test scores after the first three or so years in the classroom. This has led policymakers to pursue reforms that ignore teacher experience or seek to remove it entirely as a factor in teacher personnel policies. Our results suggest a re-evaluation of such policies.

Acknowledgments

We would like to thank the Center for Education Policy Research at the Harvard Graduate School of Education for providing the data we use in this paper. We are especially grateful to Sarah Cohodes, Jon Fullerton, Jason Grissom, Lawrence Katz, Richard Murnane, and Doug Staiger for their valuable feedback on earlier drafts. Any errors and omissions are our own.

References

Allen, J.P., Pianta, R.C., Gregory, A., Mikami, A.Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333,* 1034-1037.

Becker, G. (1993). *Human Capital*, 3rd ed. Chicago: University of Chicago Press.

Boyd, D., Lankford, H., Loeb, S., Rockoff, J., & Wyckoff, J. (2008). The narrowing gap in New York City teacher qualifications and its implications for student achievement in high-poverty schools. *Journal of Policy Analysis and Management, 27*(4), 793-818.

Carrell, S.E. & West, J.E. (2011). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy, 118*(3), 409-432.

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

Clotfelter, C.T., Ladd, H.F., & Vigdor, J.L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources, 41(4),* 778-820

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,* 673-682.

Corcoran, S.P., Evans, W.N., & Schwab, R.M. (2004). Changing labor-market opportunities for women and the quality of teachers, 1957-2000. *American Economic Review, 94*(2), 230-235.

Deaton, A. (1997). *The analysis of household surveys: A microeconometric approach to development policy*. Baltimore, Maryland: John Hopkins University Press.

Hansen, M. (2009). How career concerns influence public workers' effort: Evidence from the teacher labor market. *Urban Institute Working Paper 40*.

Hanushek, E. A., & Rivkin, S.G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*(2), 267-271.

Harris, D. & Sass, T. (2006). Value-added models and the measurement of teacher quality. Unpublished Manuscript.

Harris, D. & Sass, T. (2011). Teacher training, teacher quality, and student achievement. *Journal of Public Economics, 95,* 798-812.

Hoxby, C.M. & Leigh, A. (2004). Pulled away or pushed out? Explaining the decline of teacher aptitude in the United States. *American Economic Review, 94*(2), 236-240.

Huberman, M. (1992). Teacher development and instructional mastery. In A. Hargreaves & M. G. Fullan (Eds.), *Understand Teacher Development*. New York, Teachers College Press, pp.

122-142.

Jackson, C.K. & Bruegmann, E. (2009). Teaching students and teaching each other: the importance of peer learning for teachers. *American Economic Journal: Applied, 1*(4), 85-108.

Johnson, S.M. & the Project on the Next Generation of Teachers. (2003). *Finders and Keepers: Helping new teachers survive and thrive in our schools.* San Francisco: Jossey-Bass.

Kane, T.J., & Staiger, D.O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. NBER Working Paper No. 14607.

Kane, T.J., Rockoff, J.E., & Staiger, D.O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review, 27*, 615-631.

Rice, J.K. (2013). Learning from experience? Evidence on the impact and distribution of teacher experience and the implications for teacher policy. *Education Finance and Policy, 8*(3): 332–348.

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), 54-81.

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

Matsumura, L. C., Garnier, H. E., & Resnick, L. B. (2010). Implementing literacy coaching: The role of school social resources. *Educational Evaluation and Policy Analysis, 32*(2), 249-272.

McCaffrey, D.F., Lockwood, J.R., Koretz, D., Louis, T.A., & Hamilton, L.A. (2004). Models for value- added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*(1), 67-101

Murnane, R.J. & Phillips, B.R. (1981). Learning by doing, vintage, and selection: Three pieces of the puzzle relating teaching experience and teaching performance. *Economics of Education Review, 1*(4), 453-465.

Neuman, S. B., & Cunningham, L. (2009). The impact of professional development and coaching on early language and literacy instructional practices. *American Educational Research Journal, 46*(2), 532-566.

Ost, B. (2014). How do teachers improve? The relative importance of specific and general human capital. *American Economic Journal: Applied Economics, 6*(2): 127-51.

Papay, J.P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal,48*(1), 163-193.

Powell, D. R., Diamond, K. E., Burchinal, M. R., & Koehler, M. J. (2010). Effects of an early literacy professional development intervention on head start teachers and children. *Journal of Educational Psychology, 102*(2), 299-312.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy, 4*(4), 492–519.

Rivkin, G., Hanushek E., & Kain, J. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417-458.

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

Rockoff, J.E., Jacob, B.A., Kane, T.J. & Staiger, D.O. (2011). Can you recognize an effective teacher when you recruit one? *Education Finance & Policy, 6*(1), 43-74.

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

Scafidi, B., D.L. Sjoquist, and Todd R. Stinebrickner. 2007. Race, poverty, and teacher mobility." *Economics of Education Review, 26*:145-159.

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.

Taylor, E.S., & Tyler, J.H. (2012). The effect of evaluation on teacher performance. *American Economic Review, 102*(7), 3628-51.

TNTP. (2012). *The Irreplaceables*. New York, NY: Author.

Todd, P.E. & Wolpin, K.I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, *113*(485), F3-F3.

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

Wiswall, M. (2013). The dynamics of teacher quality. *Journal of Public Economics*. *100*, 61-78.

Figure 1. Distribution of student-year observations in mathematics estimation sample, by teacher experience level.

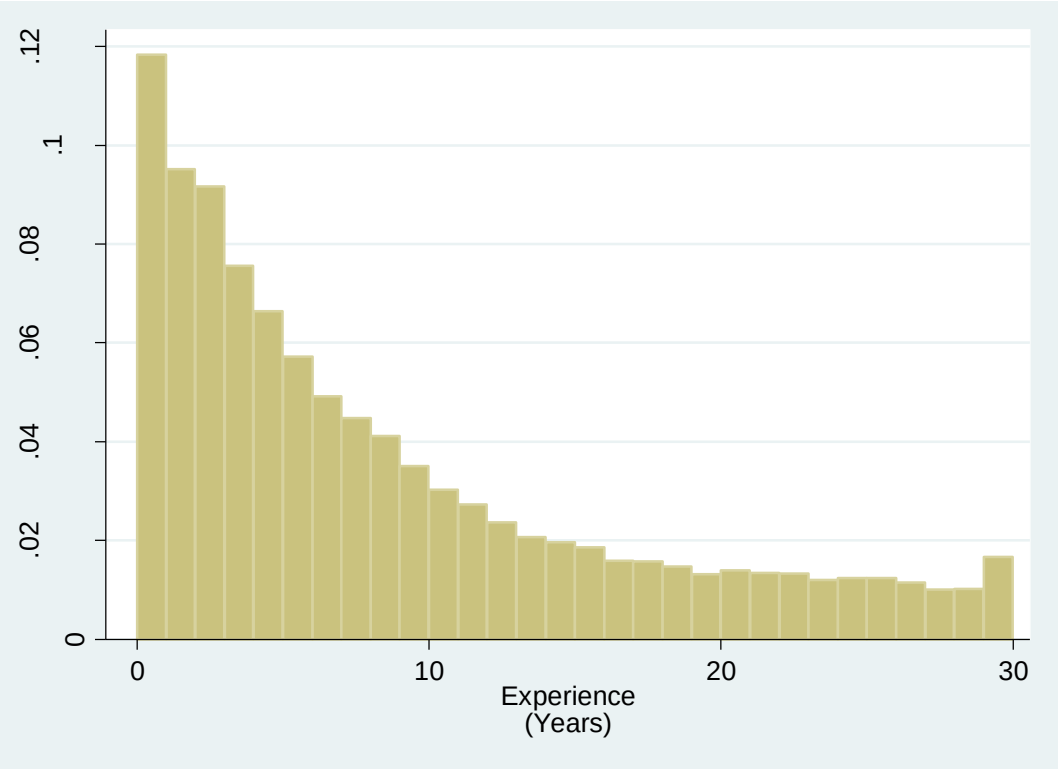


Figure 2. Series of plausible "true" productivity-experience profiles for teachers used in the simulation.

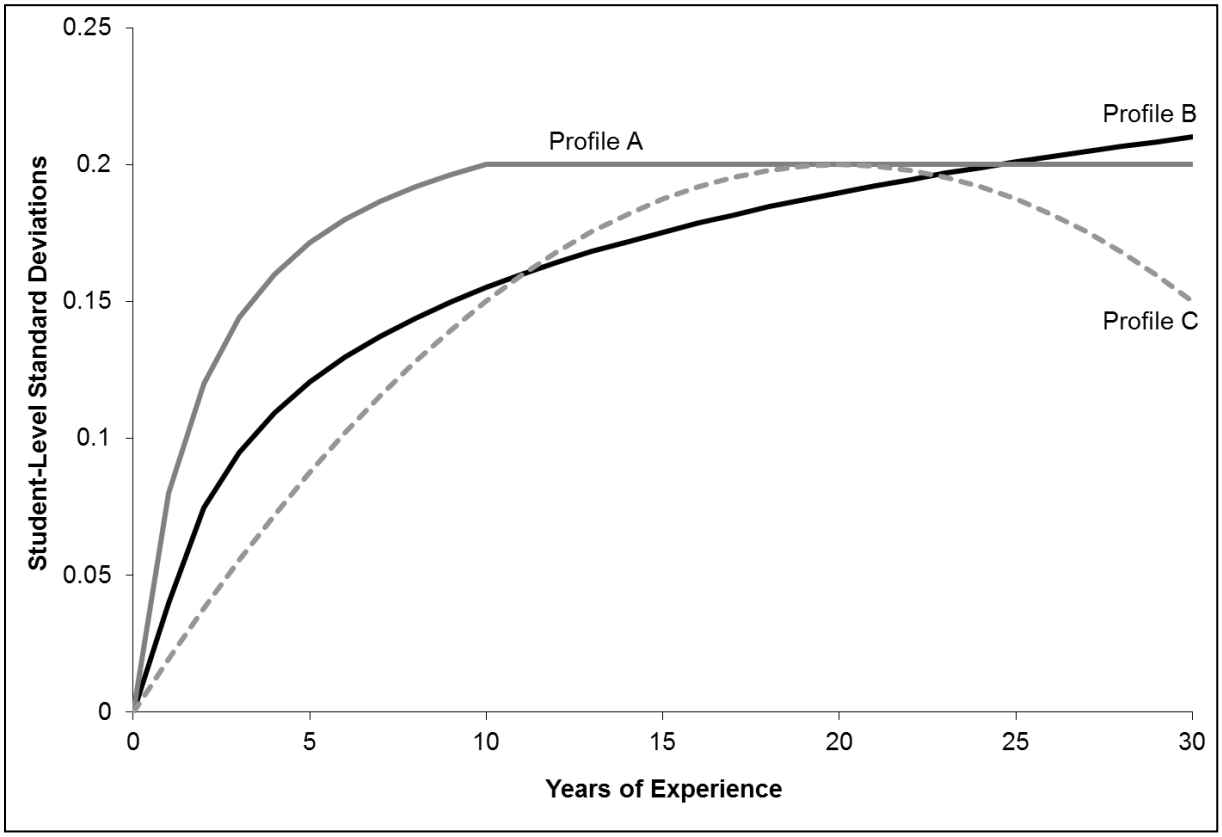
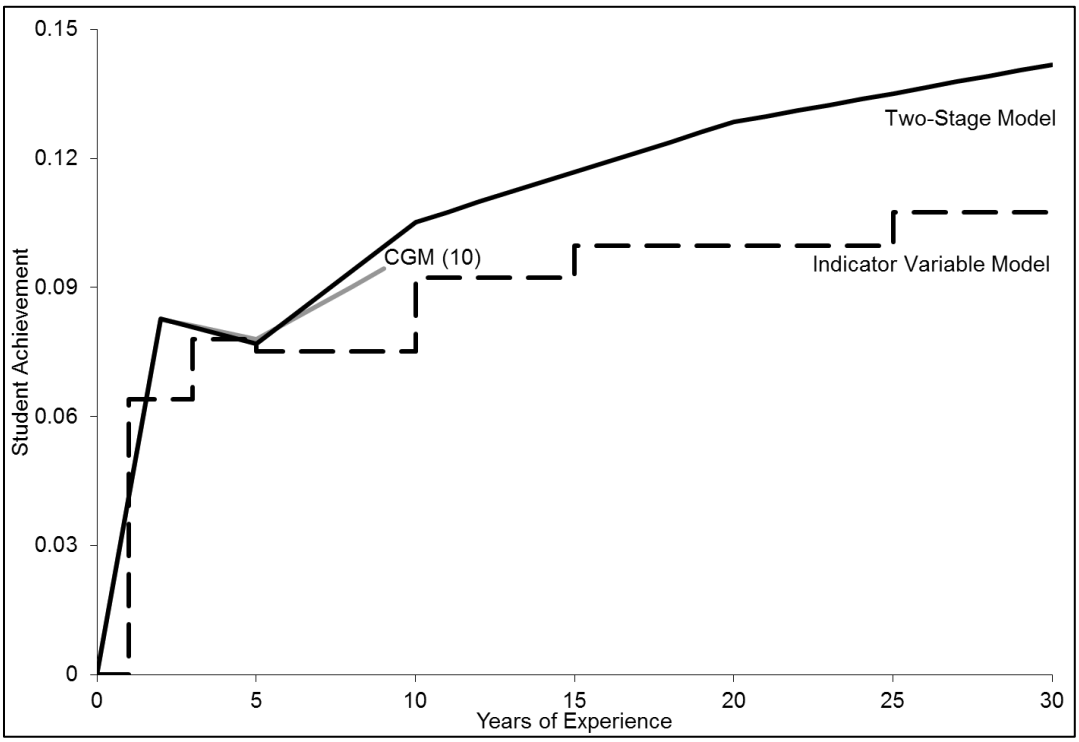


Figure 3. Estimated productivity-experience profile using the *Censored Growth Model*, the *Indicator Variable Model*, and the *Two Stage Model* in mathematics (top panel) and reading (bottom panel), from equation (4).



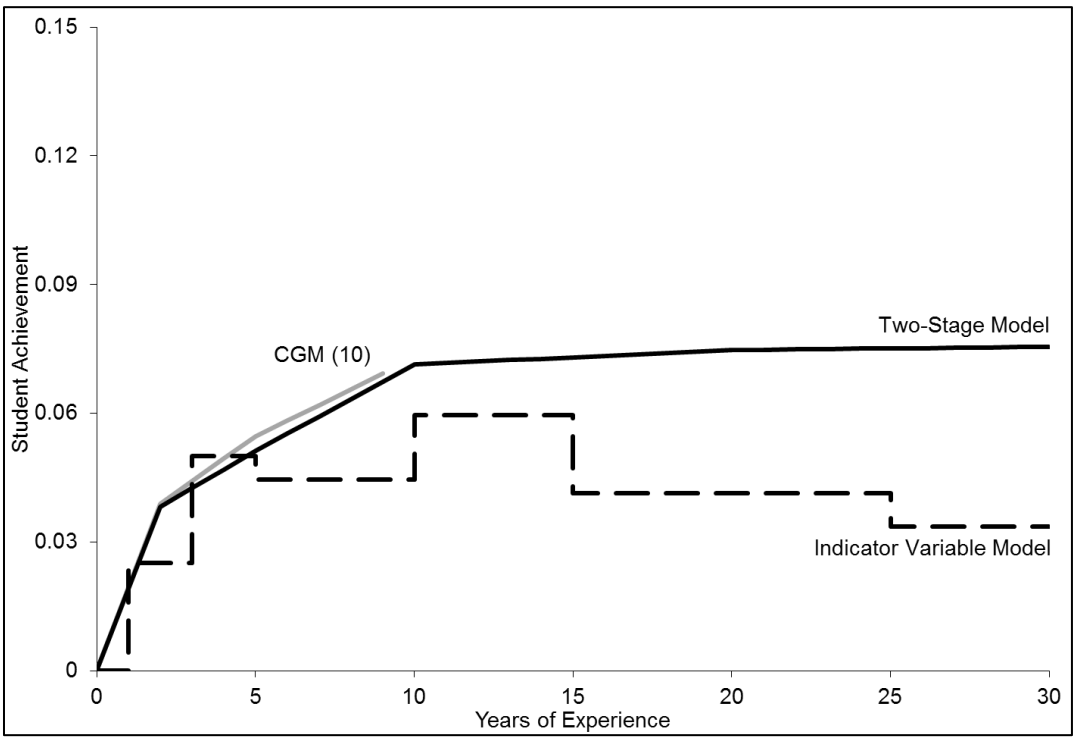


Figure 4. Estimated productivity-experience profile using the *Discontinuous Career Model*, in mathematics and reading.

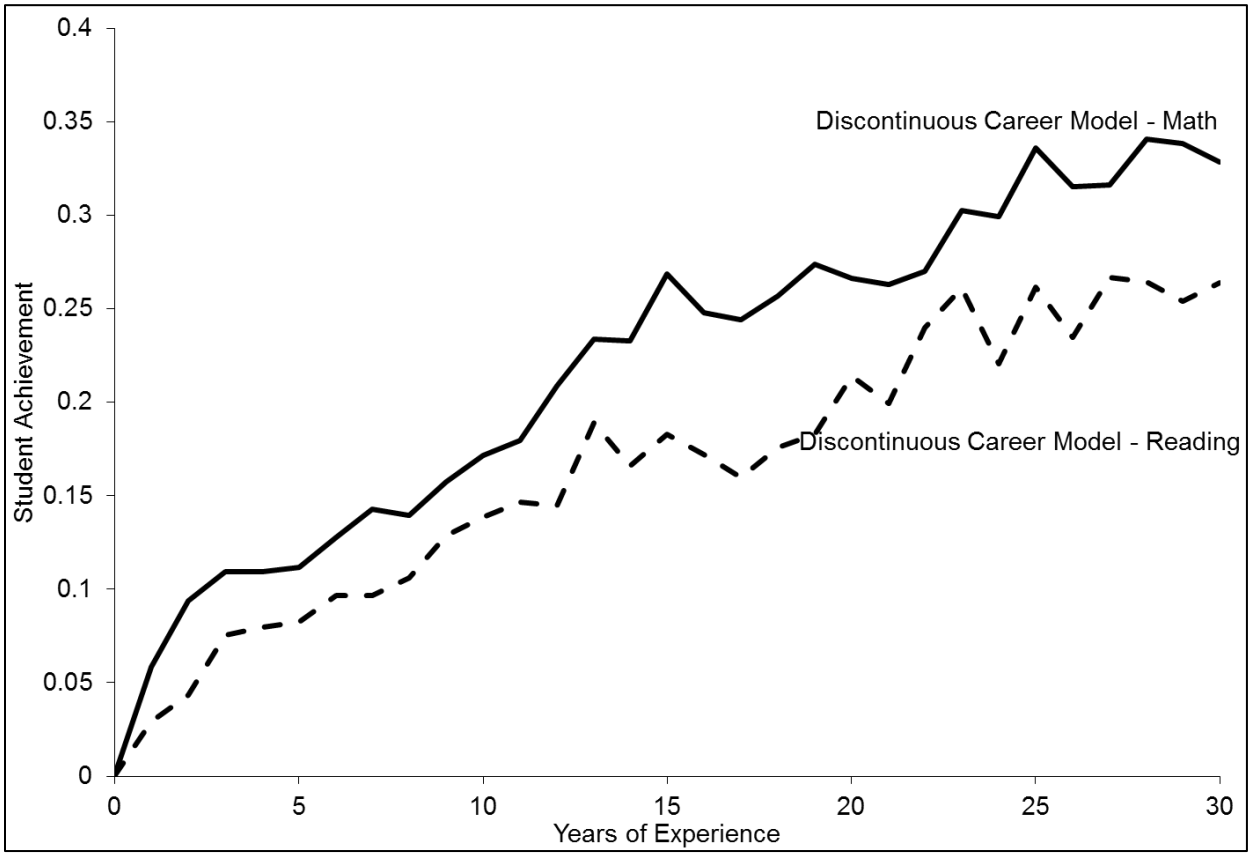


Figure 5A. Estimated productivity-experience profiles in mathematics using the *Censored Growth Model*, with cutoffs at 10, 15, and 20 years of experience.

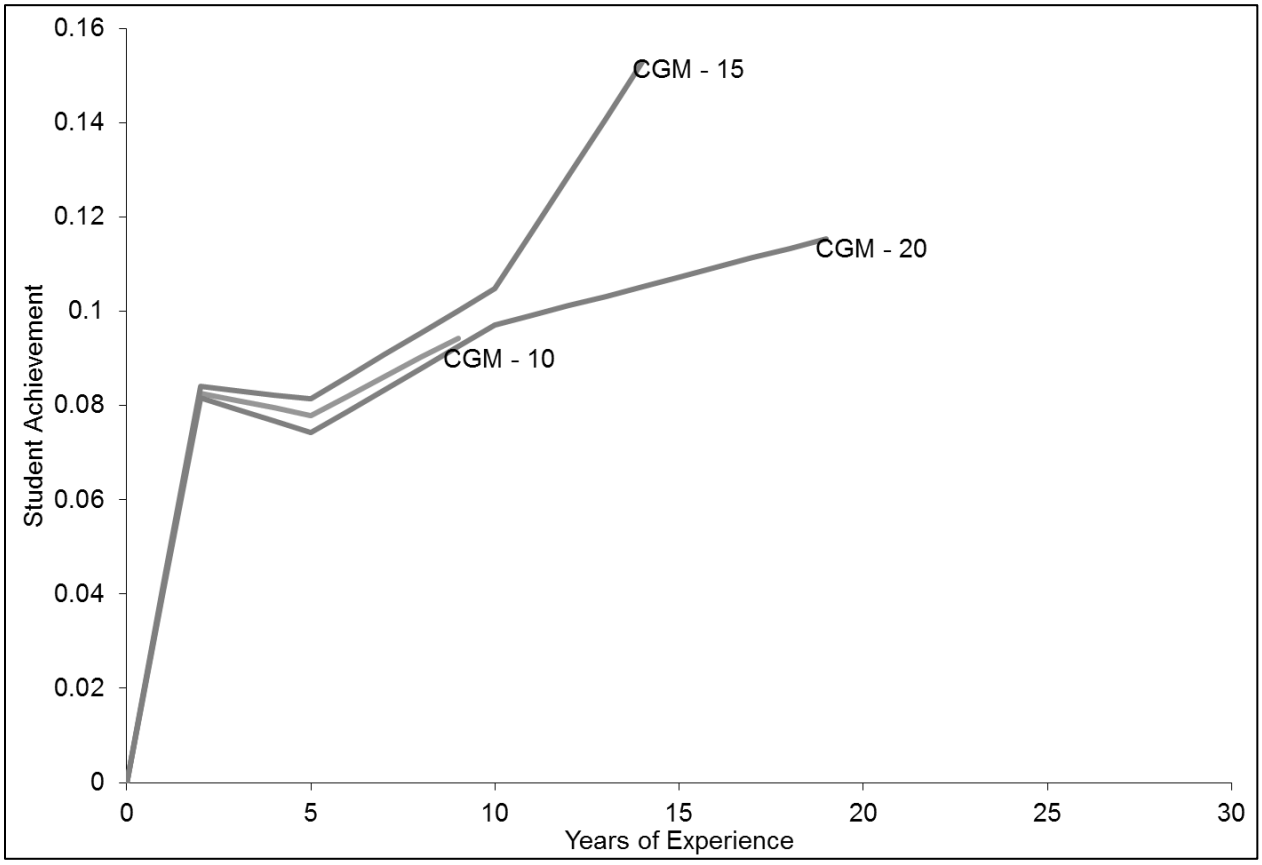


Figure 5B. Estimated productivity-experience profiles using the *Indicator Variable Model*, with different functional form specifications of experience during the first 10 years of the teaching career.

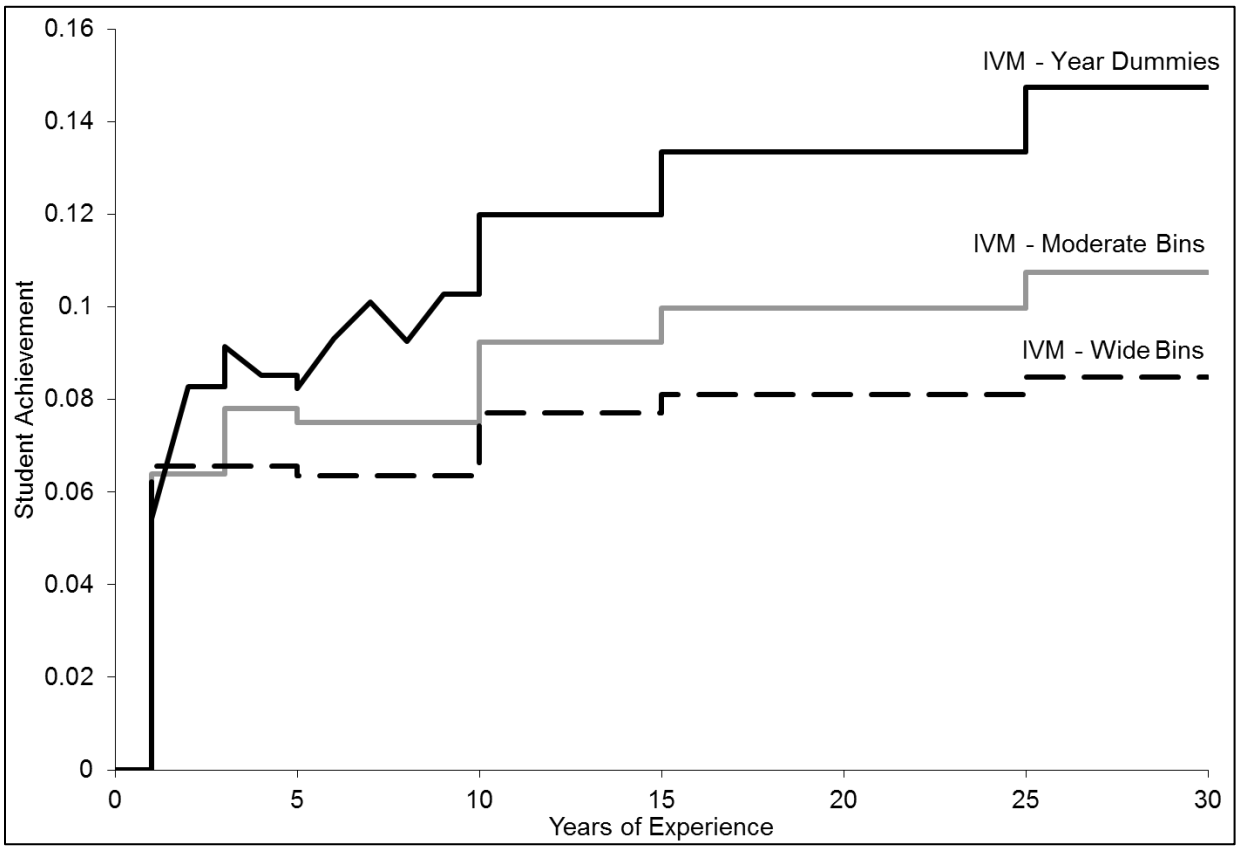


Figure 5C. Estimated productivity-experience profiles using the *Discontinuous Career Model*, from the full sample and a subsample excluding teachers who gain more than one year of teaching experience in a calendar year, in mathematics.

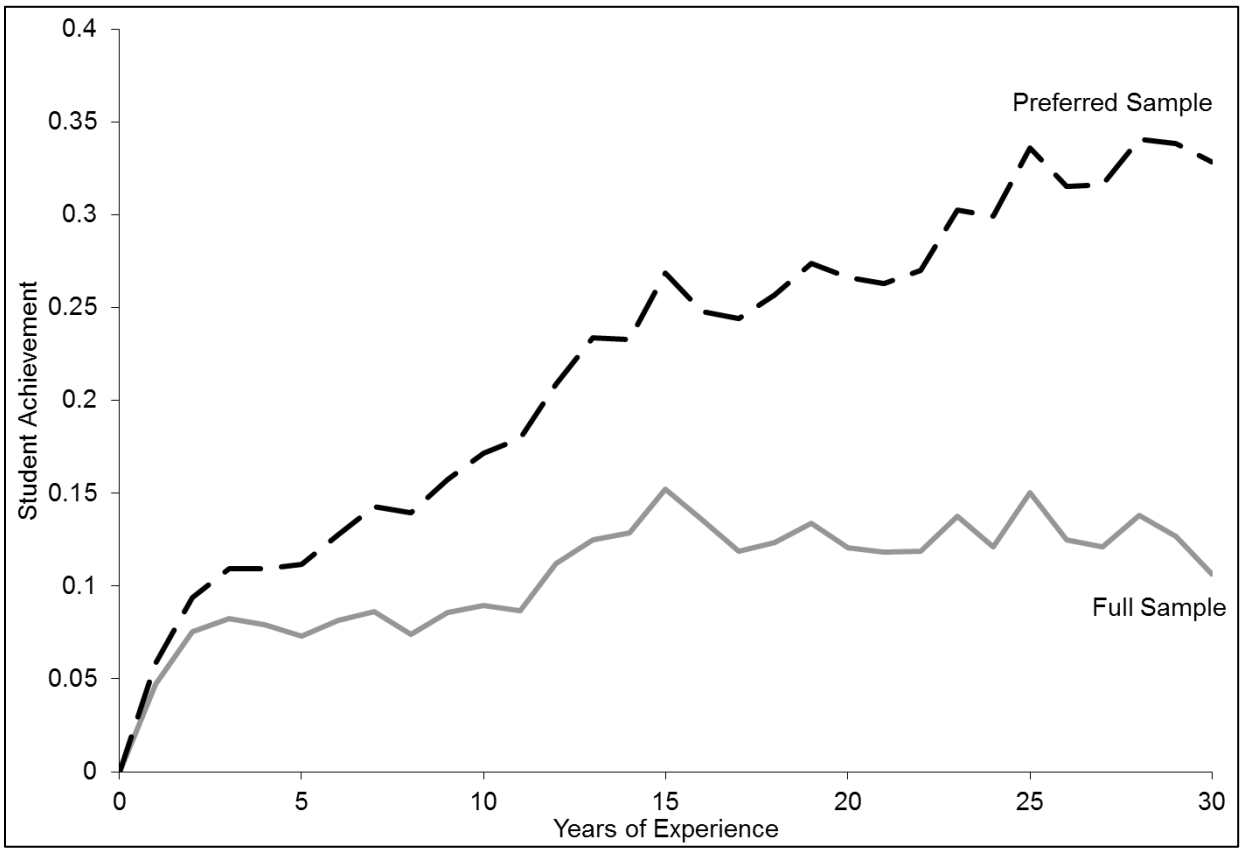


Figure 6. Estimated productivity-experience profiles using preferred versions of the *Censored Growth Model*, *Indicator Variable Model*, and *Two-Stage Model*, in mathematics (top panel) and reading (bottom panel).

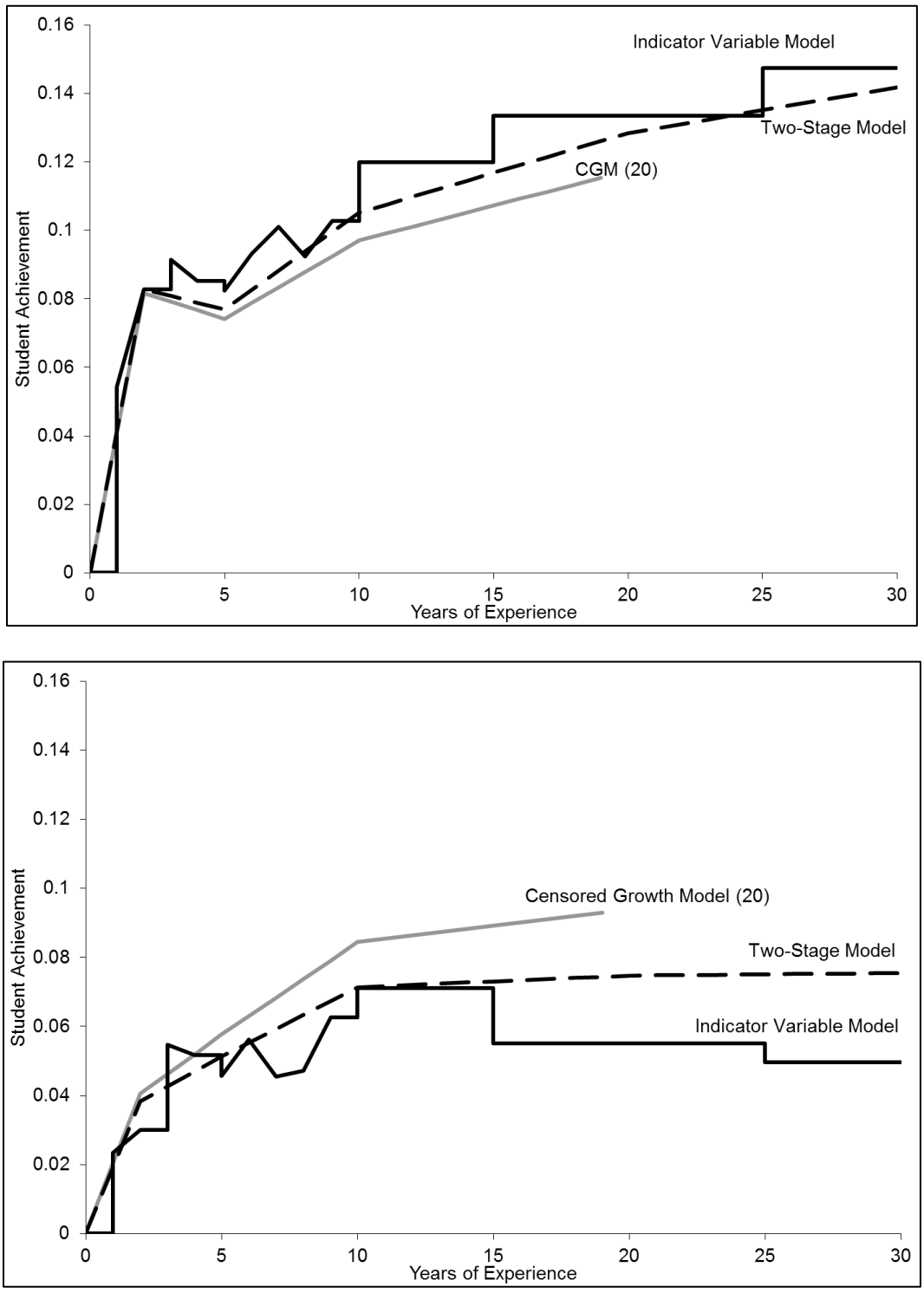


Table 1. Percent bias in implied returns to experience across models, profiles, and trends in teacher effects, at different levels of teacher experience.

Model	Profile	Teacher Fixed Effect	Years of Experience 5	10	20	30
Censored Growth Model	A	No trend	-2.1%	-1.1%	-1.1%	-1.1%
		Trend	-2.1%	-1.2%	-1.2%	-1.2%
	B	No trend	-10.2%	-16.8%	-32.0%	-38.6%
		Trend	-10.2%	-17.0%	-32.2%	-38.7%
	C	No trend	-4.2%	-3.8%	-27.8%	-3.8%
		Trend	-4.2%	-4.0%	-28.0%	-4.0%
Indicator Variable Model	A	No trend	-12.7%	-25.8%	-33.1%	-40.4%
		Trend	-12.7%	-25.9%	-33.1%	-40.3%
	B	No trend	-25.7%	-42.2%	-57.2%	-68.0%
		Trend	-25.8%	-42.3%	-57.3%	-67.9%
	C	No trend	-23.2%	-42.5%	-56.9%	-67.2%
		Trend	-23.3%	-42.6%	-56.9%	-67.1%
Two-Stage Model	A	No trend	-2.0%	-2.4%	-3.1%	-3.4%
		Trend	-8.8%	-14.1%	-26.4%	-38.3%
	B	No trend	1.4%	0.5%	0.1%	-0.5%
		Trend	-8.2%	-14.5%	-24.4%	-33.7%
	C	No trend	0.6%	2.1%	3.0%	4.0%
		Trend	-12.6%	-13.5%	-20.3%	-42.5%
Discontinuous Career Model	A	No trend	-1.5%	-1.6%	-1.4%	-0.6%
		Shock before	-9.5%	-15.2%	-28.8%	-42.9%
		Shock after	5.6%	11.0%	24.5%	39.2%
	B	No trend	1.5%	0.6%	0.2%	-0.2%
		Shock before	-9.9%	-17.0%	-28.7%	-40.4%
		Shock after	11.6%	16.7%	27.5%	37.7%
	C	No trend	1.0%	2.6%	3.9%	5.9%
		Shock before	-14.7%	-15.6%	-23.5%	-50.4%
		Shock after	14.9%	19.3%	29.8%	59.0%

NOTES: Profiles A, B, and C are shown in Figure 2. Trends in teacher effects represent a correlation of 0.05 between teacher effects and year. Temporary shocks before and after a teacher leaves are, on average, 0.025 standard deviations in the relevant year. A detailed explanation of the simulation process that produced these result is provided in Appendix A.

Table 2. Implied returns to experience across different experience ranges, by model, in mathematics (top panel) and reading (bottom panel).

	Censored Growth Model (20)	Indicator Variable Model (Dummies)	Discontinuous Career Model	Two-Stage Model
Mathematics				
0 to 5	0.0742 ***	0.0824 ***	0.1216 ***	0.0769 ***
	(0.0097)	(0.0088)	(0.0324)	(--)
	(0.0208)	(0.0184)	(0.0491)	(0.0145)
5 to 15	0.0330 ~	0.0510 ***	0.1315 *	0.0399 ~
	(0.0172)	(0.0144)	(0.0631)	(--)
	(0.0359)	(0.0274)	(0.0921)	(0.0224)
5 to 25	0.0264	0.0650 ***	0.2413 ~	0.0582
	(0.0289)	(0.0192)	(0.1264)	(--)
	(0.0578)	(0.0347)	(0.1849)	(0.0399)
10 to 25	0.0035	0.0275 *	0.1699 ~	0.0299
	(0.0219)	(0.0134)	(0.0953)	(--)
	(0.0422)	(0.0247)	(0.1397)	(0.0329)
Reading				
0 to 5	0.0576 ***	0.0457 ***	0.0824 ~	0.0512 ***
	(0.0123)	(0.0114)	(0.0471)	(--)
	(0.0175)	(0.0157)	(0.0690)	(0.0120)
5 to 15	0.0315	0.0095	0.0831	0.0218
	(0.0213)	(0.0180)	(0.0927)	(--)
	(0.0294)	(0.0240)	(0.1363)	(0.0176)
5 to 25	0.0544	0.0040	0.1513	0.0239
	(0.0358)	(0.0241)	(0.1849)	(--)
	(0.0493)	(0.0306)	(0.2736)	(0.0301)
10 to 25	0.0276	-0.0213	0.1021	0.0037
	(0.0274)	(0.0165)	(0.1390)	(--)
	(0.0366)	(0.0198)	(0.2071)	(0.0251)

NOTE: Cell entries include estimates of the returns to experience within each range of experience, traditional standard errors (except for the Two-Stage Model), bootstrapped standard errors, and approximate p-values. Results are for each of the four preferred models described in sections 4.1 and 4.3, using the Censored Growth Model with a 20 year cutoff and the Indicator Variable Model with dummies for experience levels early in the career. Results are in mathematics and reading, for the restricted sample excluding teachers with non-standard careers and the "full" sample including teachers with non-standard careers. All regressions presented exclude teachers who appear to gain more than one year of experience in a given calendar year.
~: p<0.1; *, p<0.05; **, p<0.01; ***, p<0.001

Table 3. Estimated coefficients showing the relationship between teacher attrition and past teacher productivity improvement, from equation (5), in mathematics and reading (top panel), with General Linear Hypothesis test results testing whether the relationship between attrition and past productivity improvement varies by teacher experience.

Productivity Change Measure	**Mathematics**	**Reading**
Year *t-1* to *t*	-0.0256 *	-0.0150
	(0.0126)	(0.0125)
	p=0.042	p=0.232
Year *t-2* to *t-1*	-0.0139	0.0092
	(0.0166)	(0.0179)
	p=0.401	p=0.608
Results from General Linear Hypothesis Test		
Year *t-1* to *t*	$F_{37,3562}$=0.85	$F_{37,3259}$=0.74
	p=0.733	p=0.873
Year *t-2* to *t-1*	$F_{34,2232}$=0.68	$F_{33,1980}$=0.77
	p=0.9161	p=0.827

NOTES: In panel A, each cell contains estimates, and corresponding standard errors and p-values, from a separate regression. Productivity change measures capture the difference in estimated teacher productivity (value-added to student achievement) between the prior year (*t*-1) and current year or the between two years earlier (*t*-2) and the prior year (*t*-1). In panel B, the GLH tests are on a full set of experience dummies interacted with attrition.