Task-Specific Experience and Task-Specific Talent:

Decomposing the Productivity of High School Teachers

Jason Cook Richard Mansfield

Cornell University Cornell University

PRELIMINARY AND INCOMPLETE

DO NOT CITE WITHOUT AUTHORS' PERMISSION

December 31, 2013

Abstract

In this paper, we use administrative panel data to decompose worker performance into components relating to general talent, task-specific talent, general experience, and task-specific experience. We consider the context of high school teachers, in which tasks consist of teaching particular subjects in particular tracks. Using the timing of changes in the subjects and levels to which teachers are assigned to provide identifying variation, we show that much of the productivity gains to teacher experience estimated in the literature are actually subject-specific. By contrast, very little of the variation in the permanent component of productivity among teachers is subject-specific or level-specific. *JEL Codes:* I24, I21, J45, J62.

1 Introduction

A worker's productivity at a particular job is generally assumed to depend on both the worker's innate talent as well as the worker's experience (learning by doing) performing required job tasks. Further, assessing the extent to which human capital accumulation is general to the job or specific to particular tasks has been an important research agenda for labor economists beginning with the work of Becker (1964). Knowledge of the relative importance of task-specific talent versus task-specific experience is essential for employers to maximize the productivity of their workforce. For tasks with larger potential experience gains and smaller variance in task-specific innate talent, the key to a productive workforce is employee retention: the optimal strategy is to keep employees of all talent levels at their originally assigned tasks to benefit from experience. Conversely, for tasks yielding smaller experience gains with a larger variance in task-specific talent, the optimal strategy is to fire or reassign low performing workers in an attempt to either improve general worker skill or identify superior worker-task matches.

In this paper, we use administrative panel data to decompose worker performance into components relating to general talent, task-specific talent, general experience, and task-specific experience. We consider the context of high school teachers, in which tasks consist of teaching particular subjects in particular tracks. Myriad papers have estimated education production functions featuring both teacher fixed effects and a common experience profile. The bulk of the evidence suggests that the standard deviation of permanent teacher quality is between .1 and .2 standard deviations at either the primary [1] or secondary school level[2], while teachers tend to improve with experience by around .05 test score standard deviations in their first year, another .03 to .05 over the next couple of years, and another .03 to .05 over the next several years, with the profile for mid-career teachers flattening out at between .1 and .2 standard deviations better than a novice teacher.[3]. More recent

[1]e.g. Rockoff (2004), Hanushek et al. (2005), Clotfelter et al. (2006), Harris and Sass (2006), Boyd et al. (2008), Jackson and Bruegmann (2009), Harris (2009), Harris and Sass (2011), Jackson (2013)

[2]e.g. Aaronson et al. (2007), Harris (2009), Jackson (2012), Mansfield (2013).

[3]e.g. Rivkin et al. (2005), Clotfelter et al. (2007)

studies explore the functional form assumptions utilized in such studies and find somewhat larger returns to late-career teaching.[4]

However, this literature has generally ignored the possibility that the permanent effectiveness of a teacher and/or the gains to teaching experience might be specific to a particular classroom environment. In such a context, models that impose homogeneity of skill across different classroom environments will return a weighted average of teacher skill across the environments each teacher actually faced (weighted by the fraction of time spent in each environment). To the extent that a teacher faces different classroom contexts over their career, models that impose homogeneity of returns to experience across different classroom environments may underestimate the gains to context-specific experience. Similarly, to the extent that a teacher's classroom environment has remained somewhat stable during their career, such models may overestimate the returns to general experience.

Of course, the distinction between context-specific and general skill or experience may not be critical if most teachers spend the bulk of their careers in a single context. Indeed, at the elementary and middle school levels, curriculum and classroom composition may change very little from year to year, particularly for teachers who remain in the same school. Even in this context, however, Ost (2013) shows that teachers who always repeat elementary grade assignments improve 35% faster than teachers who never repeat grade assignments.

At the high school level, however, teachers are routinely asked to teach courses in different subjects and in different difficulty tracks. Indeed, teacher certification in most states is at the level of the field (math, science, history, etc.) rather than the subject (Biology, Chemistry, Physics), and is not specific to a level of difficulty (special education excepted). If teaching skill is specific to the subject or level, then such changes in teaching assignments may have important implications for

[4]Instead of using functional form assumptions to separate total experience effects from year effects, Wiswall (2013) estimates returns to experience using discontinuous teacher careers. Papay and Kraft (2011) first model productivity as a function of both experience and year effects, omitting teacher effects. Then, in a second stage regression, they model productivity by including teacher effects and constraining year effects to be that of the first stage.

student achievement.

On one hand, suppose teachers have innate or pre-determined comparative advantages for particular subjects or difficulty levels. Then mutually advantageous swaps among teachers could produce efficiency gains if both teachers move toward their relatively more effective subjects or levels. More generally, observing a teacher in a number of different classroom contexts early in their career might produce valuable information about his/her relative teaching strengths. Permanent subject-specific skill might exist, for example, if a teacher's undergraduate major was in a particular subject (say Physics rather than Biology). Permanent level-specific skill might exist, for example, if a teacher has strong classroom control skills due to natural charisma or sense of humor, which may be comparatively more important in remedial or basic level courses, where students may tend to be less engaged.

On the other hand, suppose task-specific skill is primarily learned through experience rather than being predetermined at the time of hire. Then rotating the classroom environments to which teachers are assigned will waste a component of each teacher's skill, and slow each teacher's progress toward his/her full potential. Subject-specific experience might be important, for example, if a teacher's knowledge of the subject content deepens over time. Level-specific experience might also be significant if methods for maintaining student attention and enthusiasm depend on the level of student ability, or if the appropriate pace at which to deliver content depends on student skill and is slowly calibrated over time. In addition, experience teaching a particular subject-level combination (e.g. honors biology) might be particularly valuable if it allows teachers to hone lectures over time.

Thus, in this paper, we use administrative panel data from North Carolina to decompose effectiveness into (1) a set of permanent components capturing general talent, course-specific talent, level-specific talent, and course-level specific talent, and (2) a set of functions capturing returns from general experience, course-specific experience, level-specific experience, and course-level specific experience. The North Carolina data track teachers and students in the universe of public high schools from 1997-2009. Critically, the data feature 74,000 within-teacher changes in subject assignment and over 45,000 academic-level switches. Such rich data permit estimation of an education production function that includes general, course-specific, level-specific, and course-level-specific experience profiles as well as a full set of school-teacher-course-level fixed effects. The flexibility of our model allows us to control for many potential biases that might otherwise accompany endogenous course assignment decisions.

To preview our results, we find that a substantial portion of the return to years of experience that have been estimated in the value-added literature is actually specific to the subject that the teacher taught. We find little evidence of returns to level-specific experience and no evidence of returns to subject-level experience. [Should it be opposite?] In agreement with the rest of the value-added literature, we find that the variation in innate teaching skill is comparable in magnitude to the gains to experience; in a mild contrast to gains from experience, however, over 80% of the variance in permanent skill is general to all subjects and levels. Our estimates suggest only a minor role for subject-specific or level-specific teaching talent.

We test for and fail to find convincing evidence of estimation biases driven by dynamic assignment responses to classroom shocks, school-year shocks, or heterogeneous teacher-specific growth with experience. We do find, however, evidence of submodularity in the production function that maps general experience, subject-specific experience, and level-specific experience into teacher productivity. Specifically, gains from general and subject-specific experience accrue primarily when the teacher has relatively low levels of level-specific experience. Incorporating a richer production function that accommodates such non-separability removes a meaningful negative bias from estimates of the returns to subject-level experience in our baseline model, but does not significantly alter the qualitative conclusions outlined above.

Of course, the knowledge that a large fraction of the gains from experience are subject-specific may be of limited value to principals if most changes in course assignments are driven by necessity. For example, parental leave may require principals to reassign teachers to unfamiliar subjects or tracks.

Thus, using our estimated experience profiles, we perform a counterfactual simulation in which we assess the potential achievement gains from maximizing the context-specific skill associated with teaching assignments. Specifically, for each year of our data, we reassign the teachers observed teaching at each school in that year to the courses that were offered at their school at the time in order to maximize student performance, given the four-dimensional profiles of experience that each teacher possesses at that point in time. To ensure that our counterfactual allocation was feasible for the principal at the time, we restrict each teacher to only teach the number of classrooms they were actually observed teaching in each year.

While we can only observe the prior course and level histories of the subset of teachers who began teaching after 1995, our simulation from this subsample indicates that the statewide average gain from efficient use of context-specific experience (relative to the observed allocation) are as large as .025 student-level standard deviations by 2009, and would continue to grow over time. Since there are no changes in total teaching load for any teacher, these efficiency gains could potentially be reaped with essentially zero cost.

The rest of the paper proceeds as follows. Section 2 presents the education production function whose parameters we estimate. Section 3 describes how comparisons of teachers with different course assignment histories can provide joint identification of both teacher-course-level fixed effects and general, course-specific, level-specific, and course-level-specific experience profiles. Section 4 discusses the North Carolina administrative data and provides summary statistics displaying the variation in teacher course assignments. Section 5 considers estimation of the model. Section 6 presents the parameter estimates from our main specification. Section 7 discusses possible threats to our identifying assumptions and presents results from several specification tests and robustness checks. Notably, Section 7 demonstrates the existence of non-trivial interactions between different components of context-specific experience. Section 8 describes the counterfactual simulation in which teachers' course assignments are made to maximize gains from context-specific experience and presents the results from the simulation. Finally, Section 9 concludes.

2 Model Specification

Because our focus is on the relative importance of context-specific teacher skill and experience to test score performance, we craft our specification of the achievement production function so as to isolate the contribution of these components. Let Y_{ict} represent the standardized test score of student i in classroom c at time t. Let $r(i,c,t)$ denote the teacher that taught student i in classroom c at time t. Similarly, let $s(i,c,t)$ denote the school at which student i experienced classroom c at time t, let $j(i,c,t)$ denote the subject taught in student i's classroom c at time t, and let $l(i,c,t)$ denote the difficulty level or track associated with the classroom. $l(i,c,t) \in \{b,h\}$, where b denotes "basic" and h denotes "honors".[5] Each test score Y_{ict} is standardized so that the distribution of test scores in each subject-year combination has zero mean and unit variance.

By suppressing the dependence of s, r, j, and l on (i,c,t), we can represent the production of test score performance compactly via:

$$(1) \quad Y_{ict} = X_{ict}\beta_{jl} + \delta_{sjl} + \mu_{srjl} + d^{tot}(exp_{rt}^{tot}) + d^{j}(exp_{rt}^{j}) + d^{l}(exp_{rt}^{l}) + d^{jl}(exp_{rt}^{jl}) + \epsilon_{ict}$$

X_{ict} represents a vector of student observable characteristics and middle school reading and math test scores, along with a vector of the average levels of observable characteristics and past test scores in classroom c. We allow the impact of student and classroom characteristics and past scores to differ by subject-level combination. This allows the students' past test scores to reveal comparative advantages in particular subjects, so that a high 8th grade math score might be a stronger predictor of performance in Algebra 1 than in English 1. Similarly, classroom composition might matter more in a particular subject or level if more group work takes place in say, basic biology (labs!) than in honors math.

δ_{sjl} represents a full set of school-subject-level fixed effects. These will capture the average resid-

[5] See Section 4 for a discussion of assignment of courses to difficulty levels.

ual achievement at each school-subject-level combination, after removing the part of achievement that can be predicted based on observable student and classroom characteristics. The set of $\{\delta\}$ parameters will not only capture any school-level inputs such as principal quality, neighborhood quality, or quality of the facilities, they will also capture any variation in the quality of curricula or textbooks across subjects and levels within the school. Importantly, they will also capture the contribution of average unobserved inputs of the students who sort into particular school-subject-level combinations. Thus, the inclusion of δ_{sjl} acts as a control function that absorbs school inputs as well as any potential sorting biases that might otherwise be created by students' endogenous choices of school, subject, and level.

Rothstein (2010) documents non-random student sorting into particular classrooms across North Carolina schools. However, Kinsler (2012) retests the same data, accounting for small sample sizes, and fails to reject such non-random sorting. Additionally, while students often have choices over particular subjects and academic levels to include in their school schedules, it may be difficult for students to select into particular classes given that class assignments are often generated by automated scheduling algorithms. Still, we account for the possibility of student sorting by relying on our rich set of student covariates to absorb any sorting on the predictable component of student inputs.

μ_{rsjl} represents a full set of school-teacher-subject-level fixed effects. The average school-teacher-subject-level will be normalized to be 0 for each school-subject-level, so that μ_{rsjl} can be thought of as the deviation of a particular teacher's performance in a particular subject-level combination from the mean (student-weighted) performance of all teachers that taught at the teacher's school-subject-level combination. This specification of the contribution of teacher quality allows the estimation of a fully non-parametric joint distribution of general teacher quality and subject-specific, level-specific, and even subject-level specific comparative advantages at particular subjects, both within and across teachers. Note that by including the identity of the school in the definition of the fixed effect, we are allowing each teacher to have a completely different average skill and set of comparative advantages for particular subjects and levels at each school at which they teach (a teacher who teaches in two schools is essentially treated as two different teachers). Variation in μ_{rsjl} around a given average $\overline{\mu}_{rsl}$ will provide evidence of subject-specific skill, while variation in $\overline{\mu}_{rsjl}$ around a given average $\overline{\mu}_{rsj}$ will provide evidence of level-specific skill. One can then average the contribution of each teacher across subjects and levels $\overline{\mu}_{rs}$ and compare these averages across teachers to examine the variation in general persistent teacher quality.

exp_{rt}^{tot} represents the total years of general teaching experience that teacher r possesses at time t. $d^{tot}(*)$ is a function that captures how additional years of total experience increases a teacher's ability to improve student performance (regardless of the subjects and levels in which this experience was earned). Analogously, exp_{rt}^{j}, exp_{rt}^{l}, and exp_{rt}^{jl} represent years of experience in the relevant subject, level, and subject-level combination respectively. The $d^{j}(*)$, $d^{l}(*)$, and $d^{jl}(*)$ functions capture how additional years of subject-specific experience, level-specific experience, and subject-level specific experience increase a teacher's ability to affect student test scores. $d^{tot}(*)$, $d^{j}(*)$, $d^{l}(*)$, and $d^{jl}(*)$ are each flexibly parametrized using indicators for narrow ranges of experience.

Finally, ϵ_{ict} represents an error component which combines time varying inputs not captured by the other components of the model. In particular, we model the error component as:

$$\epsilon_{ict} = \nu_{rt} + \phi_{st} + \zeta_{ct} + e_{ict} \tag{2}$$

ν_{rt} represents year-specific deviations in a teacher's quality from what would be expected based on his/her long run skill and level of experience in the appropriate subject-level combination. ϕ_{st} captures year-specific deviations in school inputs or student sorting relative to the sample-wide average for the school-subject-level. ζ_{ct} captures classroom level shocks, such as the archetypal dog barking outside the classroom window on test day. Finally, e_{ict} represents measurement error that captures the extent to which the student's performance on the particular exam deviates from what the student could have expected to score, given his/her accumulated knowledge in the subject.

We adjust standard errors to account for the existence of each of these error components.

3 Identification

3.1 Identifying the Return to General and Context-Specific Experience

To identify the experience profiles $d^{tot}(*)$, $d^{j}(*)$, $d^{l}(*)$, and $d^{jl}(*)$, the following conditions must hold:

Assumption 1: Conditional Mean Independence of

Time-Varying Unobserved Inputs and Teacher Experience

$$E[\epsilon_{ict}|ex_{rt}^{tot} = \tilde{ex}_{rt}^{tot}, (s,r,j,l) = (\tilde{s},\tilde{r},\tilde{j},\tilde{l}), \tilde{X}_{ict}] =$$

$$E[\epsilon_{ict}|(s,r,j,l) = (\tilde{s},\tilde{r},\tilde{j},\tilde{l}), \tilde{X}_{ict}] \; \forall \, \tilde{ex} \in \mathcal{EX}, (\tilde{s},\tilde{r},\tilde{j},\tilde{l}) \in \mathcal{SRJL}, \tilde{X} \in \mathcal{X} \tag{3}$$

$$E[\epsilon_{ict}|ex_{rt}^{j} = \tilde{ex}_{rt}^{j}, (s,r,j,l) = (\tilde{s},\tilde{r},\tilde{j},\tilde{l}), \tilde{X}_{ict}] =$$

$$E[\epsilon_{ict}|(s,r,j,l) = (\tilde{s},\tilde{r},\tilde{j},\tilde{l}), \tilde{X}_{ict}] \; \forall \, \tilde{ex} \in \mathcal{EX}, (\tilde{s},\tilde{r},\tilde{j},\tilde{l}) \in \mathcal{SRJL}, \tilde{X} \in \mathcal{X} \tag{4}$$

$$E[\epsilon_{ict}|ex_{rt}^{l} = \tilde{ex}_{rt}^{l}, (s,r,j,l) = (\tilde{s},\tilde{r},\tilde{j},\tilde{l}), \tilde{X}_{ict}] =$$

$$E[\epsilon_{ict}|(s,r,j,l) = (\tilde{s},\tilde{r},\tilde{j},\tilde{l}), \tilde{X}_{ict}] \; \forall \, \tilde{ex} \in \mathcal{EX}, (\tilde{s},\tilde{r},\tilde{j},\tilde{l}) \in \mathcal{SRJL}, \tilde{X} \in \mathcal{X} \tag{5}$$

$$E[\epsilon_{ict}|ex_{rt}^{jl} = \tilde{ex}_{rt}^{jl}, (s,r,j,l) = (\tilde{s},\tilde{r},\tilde{j},\tilde{l}), \tilde{X}_{ict}] =$$

$$E[\epsilon_{ict}|(s,r,j,l) = (\tilde{s},\tilde{r},\tilde{j},\tilde{l}), \tilde{X}_{ict}] \; \forall \, \tilde{ex} \in \mathcal{EX}, (\tilde{s},\tilde{r},\tilde{j},\tilde{l}) \in \mathcal{SRJL}, \tilde{X} \in \mathcal{X} \tag{6}$$

Assumption 1 states that, for each dimension of experience, knowledge of the level of experience does not provide further information about any unobserved component of inputs, conditional on observed student inputs and the identity of the school, teacher, subject, and level. Put another way, the timing of experience accumulation in each dimension is assumed to be exogenous.

There are a number of possible threats to the validity Assumption 1, each of which relates to the exact timing of changes in experience. For example, suppose that when a school is in decline, teacher turnover begins to increase, and the teachers that remain are forced to teach both new subjects and new difficulty levels more frequently. In this case, we may observe zero subject-specific or level-specific experience more frequently when the value of ϕ_{st} is low. Since year-specific deviations in school quality from the long-run average are included in ϵ_{ict}, this scenario violates Assumption 1 and could potentially produce an overestimate of the returns to context-specific experience. Alternatively, suppose principals are reluctant to force a teacher to take on new subjects or levels when the teacher faces other short-term obstacles (such as illness, a new child, or a divorce). In that case, zero subject-specific or level-specific experience may be observed more frequently when the value of ν_{rt} is high. This scenario also violates Assumption 1, and might cause an underestimate of the returns to context-specific experience. Similarly, if teachers respond to a particularly unruly classroom by quitting teaching, or switching levels or subjects, we might underestimate the returns to experience (since those who survive to the next year of experience will have observed above-average shocks, thereby hiding the gains to the next year of experience). In Section 7, we estimate an upper bound for the degree of bias introduced by violations of Assumption 1 triggered by correlations between our experience profiles and ϕ_{st}, which we deem the most plausible of the above scenarios. We find that endogenous responses to school-year shocks are unlikely to produce a substantial bias to any of our profiles.

Despite these concerns, however, note that Assumption 1 is still much weaker than is required to identify experience profiles in most of the literature, since it conditions on the identity of the school, teacher, level, and course. Essentially, the inclusion of school-teacher-subject-level fixed effects (μ_{srcl}) controls for any arbitrary selection of teachers into experience categories based on permanent general or context-specific skill. Conditioning on r accounts for the possibility that better teachers persist long enough to gain more experience. Similarly, conditioning on r and j accounts for the possibility that the teachers allowed to gain more subject-specific experience in a particular subject are those with comparative advantages in teaching the subject, while conditioning on r and l accounts for the possibility that persistence at teaching honors level courses might signal a comparative advantage for teaching such courses.

Even if the timing of experience accumulation is conditionally independent of the error components, the simultaneous identification of each of the four experience profiles also requires considerable variation in the history of subject and level assignments across teachers.

To see concrete examples of how identification of each dimension of the experience might be secured, refer to Appendix A.

3.2 Identification of the general and context-specific components of permanent teaching skill

Identifying permanent general and context-specific teaching skill is more difficult. In particular, there is a fundamental identification problem that our model cannot overcome: we cannot distinguish average teaching quality in a school-subject-level from school or unobserved student inputs that are school-subject-level specific. If a school's students score 0.1 student level standard deviations higher in Biology than in Chemistry, we cannot tell whether all the Biology teachers are particularly effective, or if instead the Biology textbook is superior to the Chemistry textbook (or many of the student's parents are biologists). To address this issue, we consider two polar opposite assumptions, and decompose the variance in teacher permanent skill into general, subject-specific, level-specific, and subject-level specific components under each assumption. The first assumption is that average teacher effectiveness $\overline{\mu}_{rjl}$ is uniform across all levels, subjects, and schools:

Assumption 2A: Uniform Average Teacher Quality Across Contexts

(7) $$E[\overline{\mu}_{sjl}|(s,j,l)=(\tilde{s},\tilde{j},\tilde{l})]=k \text{ for some constant k}, \forall(s,j,l)\in \mathbf{RJL}$$

This would hold if the relatively more effective teachers do not sort into particular schools, subjects, or levels. Assumption 2A implies that all the variation in average residual student performance (after removing the part predictable based on student observables) across subjects, levels, and schools can be attributed to either school inputs or unobserved student inputs. Assumption 2A can be imposed on the model by including school-subject-level fixed effects (δ_{sjl}), and normalizing the student-weighted average teacher-school-subject-level fixed effect to be zero at each school-subject-level $\frac{1}{N_{sjl}} \sum_{i \in scl} \hat{\mu}_{sjl} = 0$. Under Assumption 2A, a teacher whose Biology students perform 0.1 standard deviaitons better than her Chemistry students will be assumed to be equally effective at teaching both Biology and Chemistry if the school average performance difference between Biology and Chemistry is 0.1 standard deviations. The polar opposite approach is to assume that all the variation in average residual student performance across subjects, levels, and schools can be attributed to differences in average teacher quality:

Assumption 2B: Uniform School and Unobserved Student Quality Across Contexts

$$E[\delta_{sjl} | (s, j, l) = (\tilde{s}, \tilde{j}, \tilde{l})] = k \text{ for some constant k}, \forall (s, j, l) \in \mathbf{SJL} \quad (8)$$

Assumption 2B would hold if students sort into high schools, subjects, and levels based only on observable characteristics and past performance, and all high schools and subject-level combinations within high schools provide the same contribution to student achievement. Assumption 2B can be imposed on the model by excluding school-subject-level fixed effects ($\delta_{sjl} = 0 \ \forall (s, j, l)$), and matching the between school-subject-level residual variation using a full set of teacher-school-subject-level fixed effects (without any normalizations). Under Assumption 2B, a teacher whose Biology students perform 0.1 standard deviaitons better than her Chemistry students will be assumed to be 0.1 standard deviations more effective at teaching both Biology and Chemistry if the school average performance difference between Biology and Chemistry is 0.1 standard deviations. In other words, even though the teacher is at the mean of the performance distribution in both subjects, the comparison set of Biology teachers is assumed to be 0.1 standard deviations superior on average to the comparison set of Chemistry teachers.

An intermediate assumption would be to assume that between-school variation in residual test scores is attributable to school quality and student sorting, but that the variation in residual performance that is within-schools, but across subject-level combinations is attributable to differences in average teacher quality across these combinations:

Assumption 2C: Uniform Teacher Quality Across Schools, Uniform Student/School Quality Across Subjects and Levels

$$E[\delta_{sjl}|(s,j,l)=(\tilde{s},\tilde{j},\tilde{l})] = E[\delta_{sjl}|s=\tilde{s}]\forall(s,j,l)\in\mathbf{SJL}$$

$$E[\overline{\mu}_s|s=\tilde{s}] = k \text{ for some constant } k, \forall s\in\mathbf{S} \tag{9}$$

Estimates from such a model are useful for a principal who needs to make classroom assignments for their existing stock of teachers. She may only be interested in the breakdown of within-school teacher quality into general vs. course or level-specific components, and may believe that school inputs are divided relatively equally across subjects and levels.

While Assumptions 2A-2C allow us to separate school inputs from teacher inputs, identification of $\{\mu_{srcl}\}$ also requires that other unobserved inputs are not correlated with the observation of a particular teacher in a particular course-level combination.

Assumption 3A-3C: Conditional Mean Independence of Students' Unobserved Inputs and Teacher Experience

$$E[\epsilon_{ict}|(s,r,j,l)=(\tilde{s},\tilde{r},\tilde{j},\tilde{l}),(ex^{tot}_{rt},ex^{j}_{rt},ex^{l}_{rt},ex^{jl}_{rt})=(\tilde{ex}^{tot}_{rt},\tilde{ex}^{j}_{rt},\tilde{ex}^{l}_{rt},\tilde{ex}^{jl}_{,rt}),X_{ict}=X'_{ict}]=$$

(10)

$$E[\epsilon_{ict}|(s,j,l)=(\tilde{s},\tilde{j},\tilde{l}),(ex^{tot}_{rt},ex^{j}_{rt},ex^{l}_{rt},ex^{jl}_{rt})=(\tilde{ex}^{tot}_{rt},\tilde{ex}^{j}_{rt},\tilde{ex}^{l}_{rt},\tilde{ex}^{jl}_{,rt}),X_{ict}=X'_{ict}]$$

Assumption 3B: Conditional Mean Independence of

Students' Unobserved Inputs and Teacher Experience

$$3A: \quad E[\epsilon_{ict}|(s,r,j,l)=(\tilde{s},\tilde{r},\tilde{j},\tilde{l}),(ex^{tot}_{rt},ex^{j}_{rt},ex^{l}_{rt},ex^{jl}_{rt})=(\tilde{ex}^{tot}_{rt},\tilde{ex}^{j}_{rt},\tilde{ex}^{l}_{rt},\tilde{ex}^{jl}_{,rt}),X_{ict}=X'_{ict}]=$$
$$E[\epsilon_{ict}|(s,j,l)=(\tilde{s},\tilde{j},\tilde{l}),(ex^{tot}_{rt},ex^{j}_{rt},ex^{l}_{rt},ex^{jl}_{rt})=(\tilde{ex}^{tot}_{rt},\tilde{ex}^{j}_{rt},\tilde{ex}^{l}_{rt},\tilde{ex}^{jl}_{,rt}),X_{ict}=X'_{ict}]$$
$$3B: \quad E[\epsilon_{ict}|(s,r,j,l)=(\tilde{s},\tilde{r},\tilde{j},\tilde{l}),(ex^{tot}_{rt},ex^{j}_{rt},ex^{l}_{rt},ex^{jl}_{rt})=(\tilde{ex}^{tot}_{rt},\tilde{ex}^{j}_{rt},\tilde{ex}^{l}_{rt},\tilde{ex}^{jl}_{,rt}),X_{ict}=X'_{ict}]=$$
$$E[\epsilon_{ict}|(ex^{tot}_{rt},ex^{j}_{rt},ex^{l}_{rt},ex^{jl}_{rt})=(\tilde{ex}^{tot}_{rt},\tilde{ex}^{j}_{rt},\tilde{ex}^{l}_{rt},\tilde{ex}^{jl}_{,rt}),X_{ict}=X'_{ict}]$$
$$3C: \quad E[\epsilon_{ict}|(s,r,j,l)=(\tilde{s},\tilde{r},\tilde{j},\tilde{l}),(ex^{tot}_{rt},ex^{j}_{rt},ex^{l}_{rt},ex^{jl}_{rt})=(\tilde{ex}^{tot}_{rt},\tilde{ex}^{j}_{rt},\tilde{ex}^{l}_{rt},\tilde{ex}^{jl}_{,rt}),X_{ict}=X'_{ict}]=$$

(11)

$$E[\epsilon_{ict}|s=\tilde{s},(ex^{tot}_{rt},ex^{j}_{rt},ex^{l}_{rt},ex^{jl}_{rt})=(\tilde{ex}^{tot}_{rt},\tilde{ex}^{j}_{rt},\tilde{ex}^{l}_{rt},\tilde{ex}^{jl}_{,rt}),X_{ict}=X'_{ict}]$$

Assumption 3A states that the identify of the teacher does not provide further information about any unobserved inputs, conditional on the identities of the school, subject, and track, along with the levels of both general and context-specific experience of the teacher and the observable characteristics of the student. Note that by conditioning on all four dimensions of teacher experience, we remove the concern that a teacher will be perceived to have greater general skill because they have more general experience, or more importantly, that a teacher will be perceived to have a comparative advantage at teaching in a particular context because many of the observations in that context are accompanied by considerable context-specific experience. Assumption 3B is much stronger, since it does not condition on the identity of the school, subject, or level, while Assumption 3C conditions on the identity of the school only.

Even in the case of Assumption 3A, there are still threats to the validity of Assumption 3. Suppose, for example, that a particular teacher R is assigned to a room with broken air conditioning each time she teaches honors physics (PH), but is assigned to functioning rooms whenever she teaches honors chemistry. If the teacher teaches the same number of years in each course, then conditioning on context-specific experience will not remove the correlation between the classroom-level error component ζ_{ct} and the fixed effect μ_{RPH}. Similarly, a teacher who happens to be assigned to basic English 1 (EB) classes during the years her kids are young (when she has little time to prepare for class) might exhibit a correlation between ν_{rt} and μ_{rEB}.

For a concrete example identifying how permanent teaching skill can be identified, refer to Appendix B.

Appendix B also highlights that unlike the experience profiles, each μ_{srjl} fixed effect will be estimated using only a single teacher's performance during the few years in which they taught that course-level. As such, sampling error for any given $\hat{\mu}_{rjl}$ estimate will not converge to zero even with the fairly long panel we employ. Consequently, we do not focus on individual $\hat{\mu}_{rjl}$ estimates, but seek instead to decompose the variance in performance across teachers and contexts into components attributable to general teaching talent, subject-specific talent, level-specific talent, and subject-level specific talent. Specifically, note that we can first rewrite each effect μ_{srjl} using:

$$\mu_{srjl} = \overline{\mu}_{sr} + (\mu_{srjl} - \overline{\mu}_{sr}) \tag{12}$$

The first component in 12 can be interpreted as the contribution of teacher talent that may be school-specific, but is general across subject-level combinations within the school. We will refer to $Var(\overline{\mu}_{sr})$ as the variance in general teaching talent. The second component consists of the teacher's persistent subject-level specific deviation in quality from her average level across all subject-level combinations. This can be interpreted as her comparative advantage or disadvantage at teaching subject-level combination (j, l). This second component can then be decomposed into three further components:

$$(13) \qquad (\mu_{srjl} - \overline{\mu}_{sr}) \equiv \tilde{\mu}_{srjl} = \overline{\tilde{\mu}}_{srj} + \overline{\tilde{\mu}}_{srl} + (\tilde{\mu}_{srjl} - \overline{\tilde{\mu}}_{srj} - \overline{\tilde{\mu}}_{srl})$$

The first component of 13 can be interpreted as the part of her comparative advantage at subject-level combination (j, l) that is common to all subjects. We will refer to $Var(\overline{\tilde{\mu}}_{srj})$ as the variance in subject-specific teaching talent. The second component of 13 can be interpreted as the part of her comparative advantage at subject-level combination (j, l) that is common to all levels. We will refer to $Var(\overline{\tilde{\mu}}_{srj})$ as the variance in level-specific teaching talent. The third component of 13 is the part of a teacher's comparative advantage at (j, l) that couldn't have been predicted based on the sum of her subject-specific skill and her level-specific skill. We will refer to $Var(\tilde{\mu}_{srjl} - \overline{\tilde{\mu}}_{srj} - \overline{\tilde{\mu}}_{srj})$ as the variance in subject-level specific teaching skill.

3.3 Recovering the Latent Variance Decomposition

Note that we do not observe the true variance of school-teacher-subject-level effects, but rather the sample variance, $Var(\mu_{srjl})$, which includes sampling error: $Var(\hat{\mu}_{srjl})$. To distill the true variance, we follow Aaronson et al. (2007) and Mansfield (2013). Specifically, we first define each estimated school-teacher-subject-level fixed effect $\hat{\mu}_{srjl}$ as the sum of the teacher's true context-specific skill and an uncorrelated error component: $\hat{\mu}_{srjl} = \mu_{srjl} + \xi_{srjl}$. Then the (student-weighted) sample variance in estimated context-specific skill can be decomposed as:

$$(14) \qquad \frac{1}{N} \sum_{(i,c,t) \in \mathbf{ICT}} (\hat{\mu}_{srjl})^2 = \frac{1}{N} \sum_{(i,c,t) \in \mathbf{ICT}} (\mu_{srjl})^2 + \frac{1}{N} \sum_{(i,c,t) \in \mathbf{ICT}} (\xi_{srjl})^2$$

where N is the number of test scores in the sample, and $\mathbf{ICT}$ is the set of (i,c,t) test score observations in the sample. As usual, the dependence of (s, r, j, l) on i, c, t has been dropped.

One would like to estimate the variance in true teacher quality as:

$$(15) \qquad \hat{Var}(\mu_{srjl}) = \frac{1}{N} \sum_{(i,c,t) \in \mathbf{ICT}} (\hat{\mu}_{srjl})^2 - \frac{1}{N} \sum_{(i,c,t) \in \mathbf{ICT}} (\xi_{srjl})^2.$$

ξ_{srjl} is not observed, but

$$(16) \qquad \frac{1}{N} \sum_{(i,c,t) \in \mathbf{ICT}} (\xi_{srjl})^2 \approx \frac{1}{N} \sum_{(i,c,t) \in \mathbf{ICT}} E[(\xi_{srjl})^2] = \frac{1}{N} \sum_{(i,c,t) \in \mathbf{ICT}} (se(\xi_{srjl}))^2,$$

so I estimate the error variance component using the standard error estimates for each school-teacher-subject-level fixed effect:

$$(17) \qquad \hat{Var}(\mu_{srjl}) = \frac{1}{N} \sum_{(i,c,t) \in \mathbf{ICT}} (\hat{\mu}_{srjl})^2 - \frac{1}{N} \sum_{(i,c,t) \in \mathbf{ICT}} (se(\xi_{srjl}))^2.$$

By using the delta method to estimate standard errors for $\tilde{\tilde{\mu}}_{srjl}$, denoted $se(\tilde{\tilde{\xi}}_{srjl})$, we can estimate $\hat{Var}(\tilde{\mu}_{srjl})$ analogously. Then, $\hat{Var}(\overline{\mu}_{sr})$ can be estimated via:

$$(18) \qquad \hat{Var}(\overline{\mu}_{sr}) = \hat{Var}(\mu_{srjl}) - \hat{Var}(\tilde{\mu}_{srjl})$$

To prevent teachers who only taught a single subject-level combination from biasing our estimate of $\hat{Var}(\overline{\mu}_{sr})$ downward, we restrict the sample of school-teacher-subject-level combinations when calculating $\hat{Var}(\tilde{\tilde{\mu}}_{srjl})$ to those in which the relevant school-teacher combination was observed in at least two school-teacher-subject-level combinations.

Further use of the delta method allows the same procedure to be applied in recovering the true variance of subject-specific, level-specific, and subject-level-specific teacher talent.[6]

[6]Specifically, we calculate the true variances as follows. First, consider the alternative decomposition $\tilde{\mu}_{srjl} = \overline{\tilde{\mu}}_{srj} + (\tilde{\mu}_{srjl} - \overline{\tilde{\mu}}_{srj})$. We estimate the true variance of the second component by using the delta method to calculate standard errors for $(\widehat{\tilde{\mu}_{srjl} - \overline{\tilde{\mu}}_{srj}})$ and applying the same method as above. We then obtain the variance in subject-

Finally, because we can only estimate a value of $\hat{\mu}_{srjl}$ for those combinations that we actually observe in the data, the variance in course-specific and level-specific skill that we estimate will represent the variance among the range of course and level combinations that principals actually assign. This is likely to be a selected sample; since principals may have knowledge of the relative skills of their teachers, they may avoid assigning teachers to subjects or levels at which they are likely to be particularly ineffective. For example, teaching two subjects in completely different fields (Geometry and English) may be more difficult than teaching two subjects in the same field (Algebra 1 and Geometry). While we are likely to underestimate the variance in subject-specific (or level-specific) talent across the full range of possible subjects (or levels), the estimates we do obtain may be more relevant or interesting to administrators, since nobody is proposing to assign more math teachers to English 1. The choice principals make is generally between hiring a new teacher to teach exactly the courses taught by an exiting teacher, or hiring a new teacher to teach different courses, and rotating existing teachers to new courses or levels (for example, rewarding stayers by letting them teach the honors class that was vacated by the exiting teacher). Nonetheless, we will examine whether teaching skill is actually field-specific. We can compare estimates of the variance in subject-specific skill using teachers who taught multiple subjects in the same field (e.g. Algebra 1 and Geometry) to estimates that isolate the variation in subject-specific skill derived from teachers who taught subjects in different fields (e.g. Algebra 1 and English).

4 Data

We employ data provided by the North Carolina Education Research Data Center consisting of standardized test scores of the universe of public high school students in North Carolina from 1997

specific teaching talent, $\hat{Var}(\bar{\tilde{\mu}}_{srj})$, via $\hat{Var}(\bar{\tilde{\mu}}_{srj}) = \hat{Var}(\tilde{\hat{\mu}}_{srjl}) - \hat{Var}((\tilde{\mu}_{srjl} - \bar{\tilde{\mu}}_{srj}))$. The variance in level-specific teaching talent, $\hat{Var}(\bar{\tilde{\mu}}_{srl})$, can be calculated using an identical approach. Finally, we estimate the variance in subject-level-specific teaching talent using: $\hat{Var}(\tilde{\mu}_{srjl} - \bar{\tilde{\mu}}_{srj} - \bar{\tilde{\mu}}_{srj}) = \hat{Var}(\tilde{\hat{\mu}}_{srjl}) - \hat{Var}(\bar{\tilde{\mu}}_{srj}) - \hat{Var}(\bar{\tilde{\mu}}_{srl})$.

- 2009 in eleven subjects and two class difficulty levels.[7]

During the sample period, North Carolina provided a standardized curriculum in which achievement was assessed via required end-of-course tests in English 1, Econ/Law/Politics, U.S. History, Algebra 1, Algebra 2, Geometry, Biology, Chemistry, Physics, and Physical Science.[8]

The data contain a large number of current student inputs[9] as well as past student inputs[10] and teacher-year inputs[11].

4.1 Generating the Experience Profile

We allow a flexible experience profile by creating indicators for eight experience cells: 0 years of experience, 1 year, 2 years, 3 years, 4 years, 5-6 years, 7-10 years, and 11 or more years of experience. Using this formulation, we track four types teacher experience: general experience, course-specific, difficulty-level-specific, and course-level-specific experience. Note that experience is measured in terms of years in which at least one classroom was taught in the relevant context. We posited that teaching a second classroom in the same year, when there is no opportunity to alter the lesson plan or assignments, is unlikely to provide the same experience value

[7] The data originally provide nine difficulty level delineations: Special Education, Remedial, Basic, Applied/Technical, Honors, Cooperative Education, Advanced Placement, International Baccalaureate, and non-classroom. We drop student observations coming from classes labeled as special education, cooperative education, and non-classroom. We consider remedial, basic, and applied/technical classes as "basic" and advanced placement, international baccalaureate, and honors as "honors".

[8] Testing began for Physics, Geometry, Chemistry, Physical Science, and Algebra 2 in 1999. In addition, Econ/Law/Politics was discontinued in 2004 and replaced by Civics and Economics in 2006. U.S. History was not tested between 2004 and 2005.

[9] Observable student inputs include classroom composition (including class size, racial composition, and number of gifted students in math and reading), as well as indicators for parental education, race, gender, gifted status, current or ever having Limited English Proficiency status, free/reduced price lunch eligibility, learning disability in math, reading, or writing, for whether the student intends to attend community college, attend four-year college, or work after high school, as well as indicators for participation in a sport, vocational club, academic club; service club, or arts club, and finally missing indicators for 7^{th} and 8^{th} grade math and reading scores.

[10] The student's 7^{th} and 8^{th} grade math and reading scores as well as the class's average 8^{th} grade math and reading scores

[11] Including the number of classes and number of different course-levels taught contemporaneously by the student's teacher

as teaching a classroom in a different year. However, the estimated experience profiles have very similar shapes and relative magnitudes when experience is measured using the total number of classrooms ever taught. To capture depreciation in teacher experience, we also include a set of indicators for whether the subject, level, and subject-level were taught in the previous year, as well as an indicator for whether the teacher taught at all in the previous year. We experimented with alternative specifications for capturing depreciation in experience, and the results did not change substantially. Finally, to account for possible endogenous responses to classroom shocks (explained fully in Section 6.1) we also include four indicators that equal one if the observation is from a classroom that represents the teacher's last year teaching the school-subject combination, the school-level combination, the school-subject-level combination, and at the school in any classroom, respectively.

In Table 1, we show how experience evolves for the second- and third- year teachers in our final sample. The table illustrates the variety in classroom assignment sequences. Nearly every possible pattern is represented by a significant number of teachers.

4.2 Final Specification

Our empirical strategy requires that student test score observations be matched to the teachers who taught the class. Unfortunately, the teacher ID provided in the test score data corresponds to the test administrator, who may or may not be the true teacher of the class. However, personnel records contain information on the demographic composition of each class taught by each teacher, and since the student achievement data can be aggregated to the classroom level, we utilize a fuzzy match algorithm that matches on classroom-average demographics. See Mansfield (2013) for a detailed description of the algorithm and summary statistics regarding its efficacy.

We drop from the sample test scores for students to whom we cannot match a teacher or verify a difficulty level, as well as scores from classes with fewer than 5 students. Since past test scores are critical for controlling for student sorting, we also drop observations with missing current/past test scores. Lastly, because our identification strategy relies on observing the teacher's full history of course- and level-specific experience at each point in time, we keep only test scores associated with teachers who begin teaching during our sample, as indicated by an entry level paycode.

4.3 Teacher Mobility

Table 2 depicts teacher mobility in our final sample across courses. The top entry in each cell (i, j) represents the number of teachers in our sample who ever taught in subject i that also taught in subject j, while the bottom entry represents the fraction of teachers who ever taught in subject i that also taught in subject j. The table reveals that there is considerable mobility across subjects, though the vast majority of mobility occurs within fields (e.g. math, science, etc.). This reflects the fact that certification is field-specific. Table 3 represents the corresponding transition matrix for levels. It reveals that almost all teachers who ever teach an honors class also teach at least one basic class during their career. The converse is not true; only half of teachers observed teaching at least one basic class are also observed teaching an honors class at some point during their career. This finding partly reflects the fact that there tend to be more basic courses than honors courses to staff at most schools, but is also driven by a substantial fraction of schools that do not track their classes (so that all classrooms at the school are coded as being taught at the basic level). Taken together, these tables demonstrate that teaching in multiple levels and subjects during one's career is the norm, rather than the exception.

5 Results

5.1 General and Context-Specific Experience Profiles

Table 5 displays the estimated experience profiles for each type of experience[12]. Column 1 contains estimates of the returns to teaching experience that are general to all subject-level combinations. There are considerable gains to the first two years of general experience, such that teachers teaching in their third year can expect to improve student performance by .071 test score standard deviations more than a novice teacher, even if they are teaching at a new level in a new subject. These gains persist, but are not compounded by additional years of general experience. The results become quite noisy for higher levels of experience; since we must observe the entire history of teacher assignments, only the cohorts of new teachers from the late 1990's are observed at the higher levels of experience in our sample. The magnitudes of these estimates are somewhat smaller than the standard returns to experience estimated in the literature. The results in Column 2 shed light on the source of this discrepancy: a substantial portion of the returns to experience generally estimated in the literature are actually specific to the subject the teacher taught. Since teachers frequently reteach the same subject many times, subject-specific experience and total years of experience are highly correlated. Thus, when returns to context-specific experience are not separated from returns to overall years of experience, the returns to subject-specific experience will generally be reflected in a larger estimated returns to general experience.

Column 2 also shows that teaching a subject for the second time increases the teacher's expected performance by .023 test-score standard deviations, relative to the first attempt. An extra year of subject-specific experience increases performance by an additional .024 standard deviations, while a third year of subject-experience adds an additional .023 standard deviations. Gains seem to slow beyond the third year of subject experience, but do not fully level off. Overall, teachers with

[12]Experience is ex-ante, e.g. a teacher starting their second teaching will be counted as having one year of experience.

more than 7 years of subject-specific experience are between .08 and .09 student level standard deviations more effective than teachers with the same total years of general teaching experience but who are teaching the subject for the first time.

Columns 3 and 4, by contrast, show that the returns to level-specific and subject-level-specific experience seem to be virtually non-existent, once years of course-specific and total experience have been taken into account. In fact, the returns to level-specific and subject-level specific experience seem to be negative. In Section 8, we present evidence suggesting that these negative estimates are spurious, and result from incorrectly imposing that the separate components of experience are additively separable in the education production function. Even with a more general specification, however, we find relatively small returns to level- and subject-level-specific experience.

Column 5 in Table 1 sums across the first four columns to provide the returns to experience for a teacher who never changes the subject-level he/she teaches. After 4 years, such a teacher is predicted to perform 0.11 standard deviations better than a novice teacher. Since many teachers teach the same course-level every year (perhaps in addition to other courses), this sum is particularly well identified. Most of the sampling error in the estimates comes from decomposing this sum into the four experience components.

Given the failure to observe meaningful level-specific and subject-level-specific experience effects, in the first two columns of Table 6, we present results from a specification in which all elements of the level-specific and subject-level specific experience profiles are restricted to be zero. The basic pattern of results for total and subject-specific experience do not change much; there are still significant gains from the first two years of total and subject-specific experience, and these gains generally seem to persist, but do not further accumulate. However, the magnitudes of the estimates are only about 2/3 as large as the baseline specification. Imposing the restrictions increases the precision of the estimates considerably, however, so that experienced teachers are still significantly more effective than novice teachers at nearly all experience levels for either general or subject-specific experience.

The third column of Table 6 presents estimates from the standard specification in the literature, in which only general experience enters the production function. We see that this experience profile matches fairly closely those found in the literature,[13] suggesting that the smaller returns to general experience are indeed driven by properly accounting for context-specific experience, rather than the focus on high school teachers versus elementary or middle school teachers.

Finally, both our baseline specification and restricted specification impose that the general and subject-specific experience are the same across fields. In Table 7, we present separate general and subject-specific experience profiles for math, science, social studies, and English subjects. Comparing the first six columns, we see that general and subject-specific returns to experience are quite similar across math, science, and social studies, providing support for the pooled specifications above. However, only the first year of general experience seems to have any value for English teachers, and we find no evidence of any returns to re-teaching English 1, relative to general experience teaching other English classes. This coincides with Mansfield (2013), who finds that the variance in productivity among teachers is quite stable across 9 of the 10 tested subjects, but that the variance in productivity among English teachers is only half as large. This evidence suggests that perhaps the English 1 exam does a particularly poor job of capturing the contributions of teachers. In light of these results, we do not include English classrooms in our simulations of efficiency gains from optimal assignment of teachers to classrooms presented in Section 9.

5.2 The Variance of General and Context-Specific Components of Permanent Teaching Talent

Table 8 contains the results of the decomposition of the variance in permanent teacher quality into general, subject-specific, level-specific, and subject-level specific components. The first column displays the decomposition obtained from imposing Assumption 2A, in which all between school-

[13]See Figure 1 of Atteberry et al. (2013) for a synthesis of the literature.

subject-level variation in student performance is attributed to differences in school and unobserved student inputs. The row labeled “School-Subject-Level-Teacher FE” provides the total estimated variance in teaching effectiveness across randomly sampled school-teacher-subject-level combinations, which combines all four permanent components. The point estimate is .0145, implying that a one standard deviation increase in combined permanent teaching effectiveness is associated with a .120 standard deviation increase in expected student performance. 79% of this variance in permanent teacher quality can be attributed to general teacher talent that is common to all subject-level combinations (See the row labeled “School-Teacher FE”). A student assigned to a teacher whose average effectiveness across the subject-level combinations he/she teaches is one standard deviation above average can expect a .107 standard deviation increase in test score performance relative to being assigned the average teacher at the school in the absence of knowledge about the chosen teacher’s level-specific or subject-specific skill.

Subject-specific skill and level-specific skill each make up about 13% and 6% of the total variance in permanent teaching effectiveness across randomly chosen school-teacher-subject-level combinations. Getting a teacher whose subject-specific skill is one standard deviation above the average for a particular subject increases expected performance by about .054 test score standard deviations. Note that this is still enough to move a student who would have otherwise scored at the 50th percentile to the 52nd percentile statewide. However, the variation in permanent subject-specific skill is quite small relative to the returns to subject-specific experience discussed above. Getting a teacher whose level-specific skill is one standard deviation above the average for a particular level increases expected performance by .030 test score standard deviations, only enough to move a student from the 50th to the 51st percentile.

Finally, the subject-specific, level-specific, and general components of permanent skill combine to explain nearly the full variance in permanent teacher skill across classroom contexts. There does not seem to be such a thing as subject-level-specific talent. In other words, a teacher’s permanent talent for teaching, say, honors biology, can be fully explained by the teacher’s general teaching talent across subjects and levels, combined with his/her talent for teaching honors level subjects and his/her talent for teaching biology courses, respectively.

Column 3 of Table 8 shows the alternative decomposition of permanent teacher skill that comes from imposing Assumption 2B, in which all variation in average student performance across school-subject-level combinations is attributed to differences in average teacher quality. Not surprisingly, this increases each of the variance components substantially. Note, though, that the fractions of variance explained by each component stay roughly similar to what they were under Assumption 2A. Perhaps the most compelling result from Column 5 is that the variance in level-specific skill is still only .0012, even under an assumption designed to maximize the variation attributed to teacher talent. Similarly, subject-level-specific talent does not appear to exist under Assumption 2B either. Under Assumption 2B, a one standard deviation increase in general teacher talent is associated with a .201 increase in average student performance across subject-level combinations, while a one standard deviation increase in course-specific teacher talent is associated with a .067 increase in expected student performance relative to a teacher with no comparative advantage or disadvantage at teaching the chosen subject.

The results under Assumption 2C (Column 2) stem from removing only the between-school variation from the component attributed to general teacher talent. They provide a middle ground estimate of the standard deviation in general teacher talent of .166 test score standard deviations. These results are roughly in line with those of Mansfield (2013).

6 Testing the Validity of the Identifying Assumptions

6.1 Controlling for Endogenous Responses to Classroom Shocks

A potential threat to the validity of the experience profile estimates raised in Section 3 is generated by the possibility that teachers future classroom assignments, or their willingness to continue teaching more generally, is driven by the classroom-level shocks they have received. For example, a teacher may be so vexed by a group of particularly troublesome students that they quit teaching or switch schools. Similarly, assignment to a particular poor physical classroom for Biology that undermines student learning may cause the teacher to advocate for a switch to Chemistry. These scenarios would imply that the set of teachers who make it to the next year of teaching (or course-specific teaching) are those whose classroom shocks were not too negative. Assuming classroom shocks are serially uncorrelated, the expected change in classroom shock would be negative among those who persist. Since experience profiles are entirely identified by the growth in teacher performance from one experience category to the next, such scenarios could produce underestimates of the returns to experience. We address this possibility by adding four indicator variables that are set to one if the observation is from a classroom that represents the teacher's last year teaching the school-subject combination, the school-level combination, the school-subject-level combination, and at the school in any classroom, respectively. While these dummy variables control for the most plausible dynamic response to classroom shocks, they also control for the possibility that teachers who anticipate quitting try less hard in their final year (which could also bias downward the estimated experience profile).

6.2 Testing for Endogenous Responses to School-Year Shocks

A second potential bias in the experience profile estimates stems from the possibility that reallocation of teachers across subjects and levels might be more likely when a school is enduring its relatively ineffective years (independently of the contributions of its teachers). This could occur if an inexperienced principal enters the school who has a different conception of how teachers should be allocated. It could also occur if teachers are more likely to quit during a school's relatively ineffective years, creating holes in course or level offerings that other teachers must be forced to fill. One way to test for this possibility is to examine whether schools' relatively low (or relatively high) year-specific residuals disproportionately occur with particular experience profiles. However, if we use residuals from the estimated model, any correlation between experience profiles and school-year deviations will already be reflected in biased experience profile estimates, so that the residual will have been purged of any information it might have contained about endogenous responses to school-year shocks. On the other hand, if we use residuals in which the estimated experience profiles have not been removed, then school-year average residuals will naturally be correlated with the experience profile composition of the teachers in the school-year via the causal effect of teacher context-specific experience.

The second problem can be solved, however, by re-weighting the classroom residual averages that compose school-year averages to account for differences in experience profile composition across school-years. To see the mechanics of how this might be done, refer to Appendix C.1. Intuitively, consider the residual test score where everything from the baseline model has been removed except for experience profiles.

We show that schools which have a disproportionate fraction of classrooms taught by teachers with low stock of context-specific experience will tend to have negative school-year average residuals. Instead of weighting by the fraction of classrooms in a particular school-year, this effect of experience can be removed by reweighting by the fraction of classrooms in the full sample featuring teachers with the same experience profile as the teacher who taught in the given classroom.

Given these reweighted school-year shocks, we can then examine whether particular experience levels disproportionately occur during schools' relatively ineffective years by taking a profile-specific weighted average of these shocks, where the weight on each school-year is the fraction of classrooms in the full sample featuring the experience profile in question that appeared within that school-year.

After forming the set of estimated school-year shocks ($\{\tilde{Z}_{st}\}$ from Appendix C.1), we calculate averages of these shocks for each four dimensional experience profile, then regress these averages on the design matrix for the four additively separable experience profiles to ascertain the degree to which the bias will be reflected in each of the four dimensions.

Table 9 displays these upper bound estimates of bias from endogenous responses to school-year shocks. The estimates are less than .01 student-level standard deviations for nearly all levels of experience across the four profiles. While there may be a slight downward bias in the returns to general experience and a slight upward bias in the returns to subject-specific experience, the magnitudes are far too small to explain the general pattern of results. Furthermore, as emphasized, the true bias from endogenous responses to school-year shocks is likely to be even smaller, given that the test itself is biased in the direction of the estimates of returns from Table 5.

6.3 Misspecification Tests: Testing for Classroom Assignment based on Heterogeneous Teacher Growth

A third possible violation of Assumption 1 could arise if particular experience profiles are more likely to be observed during years in which teachers are experiencing positive or negative year-specific deviations in quality relative to their predicted quality based on their performance in the full sample and their observed levels of each dimension of experience.

While there are a variety of scenarios that could bring about such a correlation[14], one particularly plausible mechanism stems from the possibility of heterogeneity in the gains to general experience among teachers. Atteberry et al. (2013) explore how different the returns to experience are in

[14]For example, teachers who divorce (and would otherwise have below-average productivity during the process year) may be less likely to request or be assigned new subjects or levels while coping with the problems at home.

the first 5 years across different levels of value added and find evidence of heterogenous teacher growth. They categorize teachers into 5 quintiles of VA measures based on the first year teaching as well as after 5 years of teaching. They find that 62% of the lowest quintile math teachers in their first year of teaching are still in the bottom 2 quintiles in year 5 and 73% of the initially highest performing teachers remain in the top two quintiles in year 5. There is also a sizable fraction of teachers switching extremes of the value added distribution after 5 years. "About 19% of bottom- and 10% of top- quintile teachers end up in the opposite extreme two quintiles," (Atteberry et al. (2013)).

Since the main specification in equation 1 constrains the gains from general experience to be common to all teachers, any heterogeneity in rates of growth among teachers in the sample will be reflected in the teacher-year error component, ν_{rt}. If assignments to particular subjects or levels are nearly evenly distributed between the later and earlier years of the teacher's career within the sample, a given teacher's value of ν_{rt} would not predict the identity of the subject-level, so that estimates of μ_{srjl} would be unbiased. Furthermore, to the extent that assignments to particular subjects or levels are not evenly distributed, such heterogeneity would generally tend to inflate the estimated variances of the variance components of context-specific permanent talent (since we would see spurious within-teacher variation in student performance across the teacher's subject-level combinations).[15] Given that the estimated true variances of subject-specific, level-specific, and subject-level specific permanent talent are all fairly trivial, we ignore the possibility of bias in our estimates of context-specific permanent talent.[16]

However, our context-specific experience profiles could be biased upward if teachers with faster than average growth rates are more likely to stay in the courses and levels they are teaching, since the average value of ν_{rt} would be higher for higher values of course-specific or level-specific experience. One plausible scenario in which this could occur is one in which rapidly improving

[15]This would be true unless the teachers who were improving faster than the sample average were systematically moving to their relatively ineffective subjects, which does not seem particularly plausible.

[16]Recall that ν_{rt} is orthogonal to general permanent talent $\overline{\mu}_r$ by construction, since ν_{rt} reflects deviations from a teacher's average performance over the full sample.

teachers are rewarded for their students performance by getting to keep their subjects and levels (while forcing others to adjust to changing classroom demand created by, say, teacher turnover or variation in student cohort size). We test this hypothesis by examining whether the trend in a teacher's performance predicts the teacher's future teaching assignments.

Specifically, let $\overline{Z}_{srjlt}$ represent the average test score residuals of students taught by teacher r in school s in subject j and level l in year t:

$$\overline{Z}_{srjlt} = \frac{1}{N_{srjlt}} \sum_{ict \in srjlt} Y_{ict} - X_{ict}\hat{\beta}_{jl} - \hat{\delta}_{sjl} - \hat{\mu}_{srjl} - \hat{f}(exp^{tot}_{rt}, exp^{j}_{rjt}, exp^{l}_{rlt}, exp^{jl}_{rjlt})$$

We estimate specifications of the following form:

$$1(AssignmentChange) = \beta_0 exp^{jl}_{srjlt} + \beta_1(\overline{Z}_{srjlt''} - \overline{Z}_{srjlt'}) + \epsilon_{stcly} \tag{19}$$

To operationalize these specifications, three choices must be made.

First, the future period over which the pre-existing trend is allowed to affect assignment patterns must be specified. We consider two choices: whether the assignment is repeated in the following year, and whether the assignment is ever repeated.

Second, the past period over which the pre-existing trend is allowed to affect assignment patterns must be specified. We consider several possibilities: the first two, three, four, five, or six years of (general or context-specific) teaching (denoted “1-2”,“1-3”,etc. in Table 10 below) as well as the most recent two years, and the most recent four years (denoted “2yr pooled” and “4yr pooled” in Table 10 below).

Third the context-specificity of the pre-existing trend must be specified. We consider four choices: the general trend in average student residual performance across all subjects and levels, the performance trend within a particular subject, the trend within a particular level, and the trend within a particular subject-level combination. The definition of an assignment change then matches the choice of context-specificity. If we consider a general performance trend, the outcome variable (denoted $1(Change)$ below) will either be an indicator for whether the teacher taught any classroom in the chosen school in the following year (denoted “Next Year” in Table 10), or an indicator for whether the teacher taught any classroom in the chosen school ever again during the sample period (denoted “Ever” in Table 10). If we consider a subject-specific trend, $1(Change)$ will indicate whether the teacher taught the same subject the next year (or, alternatively, in any future year during the sample. Level-specific and subject-level specific assignment change indicators are defined analogously.

The specification is estimated as a linear probability model, and the parameter of interest in each specification is β_1, which can be interpreted as the increase in the probability of an assignment change per 1 test-score level standard deviation increase in the existing trend.

Table 10 displays the value of $\hat{\beta}_1$ for each specification of equation 19. No clear pattern of dynamic assignment emerges for any dimension of context specificity. While a handful of the specifications report statistically significant coefficient values, the signs of the coefficients are split between positives and negatives, even within each dimension of context-specificity. Furthermore, the magnitudes of the coefficients are quite small. Note that even a large trend in teacher performance would consist of an improvement of .1 standard deviations per year. Thus, a coefficient of -.1 (among the largest of those observed) indicates that a teacher whose average student residual is increasing by .1 standard deviation per year over the chosen prior period is 1 percent less likely to change assignments the following year (or ever, depending on the specification). Thus, we find very little evidence that teacher classroom assignments (or teacher quits) are dynamically chosen on the basis of heterogeneous teacher-specific growth rates or short-term teacher-specific productivity shocks.

6.4 Misspecification Tests: Testing the Additive Separability of Context-Specific Experience Profiles

Another form of misspecification bias could arise from the restriction in equation 1 that the effect of the four dimensions of general and context-specific experience can be represented as the sum of four additively separable dimension-specific experience profiles: $d(exp^{tot}, exp^{j}, exp^{l}, exp^{jl}) = d^{tot}(exp^{tot}) + d^{j}(exp^{j}) + d^{l}(exp^{l}) + d^{jl}(exp^{jl})$. However, general experience and different dimensions of context-specific experience may interact with one another. For example, perhaps students only learn if the teacher has developed effective ways to both explain a subject's content and maintain control of the classroom. Lectures that deliver content effectively may require subject-specific experience, whereas classroom control skills may be learned through general or level-specific experience. Alternatively, perhaps a teacher can keep student attention by either having exceptional command of the content or by having excellent classroom control skills, making the production function submodular.

We relax the additive separability assumption by allowing the experience contribution to productivity to be captured by a non-parametric function of the four components of teacher experience:

$$Y_{ict} = X_{ict}\beta_{jl} + \delta_{sjl} + \mu_{srjl} + d(exp^{tot}, exp^{j}, exp^{l}, exp^{jl}) + \epsilon_{ict} \tag{20}$$

We implement this specification by replacing the four dimension-specific experience profiles with a full set of four-dimensional experience cell fixed effects. This specification is isomorphic in structure to a model with worker and firm fixed effects. Thus, satisfying the rank condition for identification in this specification requires that four-dimensional experience cells and school-teacher-subject-level cells form a connected graph, with the experience cells as vertices and school-teacher-subject-level cells as edges (or vice versa). The largest connected component of the full graph in our sample includes 347 out of the 377 observed four-dimensional experience cells and 747,890 out of the 852,115 observations that satisfy our other sample restrictions. In practice, we actually estimate all of our specifications using this restricted sample for ease of comparison and to make sure that our identification of school-teacher-subject-level effects comes primarily from within-experience cell variation.[17]

Not surprisingly, the estimated experience cell fixed effects are measured with considerable sampling error, making it difficult to distill potential complementarities.

For a full technical description of how we deal with sampling error, refer to Appendix C.2.1. Intuitively, we smooth our estimates using a kernel based on an L1 norm weighted[18] distance between experience profiles.

With a four-dimensional function, there are more potential complementarities than we can possibly test for. One approach, adopted in Figures 1 - 3 is to plot the gains to increasing values of a single dimension of experience while conditioning on different subsets of the other three dimensions of experience. In nearly every figure, we see that the gains to additional years of experience in one dimension are decreasing in the stock of experience in the second dimension, suggesting that the true production function is submodular; alternative components of experience seem to be strong substitutes for one other. While the smoothing process itself does bring the various plotted profiles closer to one another (since they are essentially differently-weighted averages of the same $\hat{d}$ profiles[19]), even $\tilde{d}$ profiles with quite different weights on the estimated $\hat{d}$ seem to exhibit similar same shapes and levels.

An alternative approach to assessing interactions in the production function is to calculate cross-partial derivatives. While our production function is estimated at only a finite number of four-dimensional combinations, there are nonetheless several combinations of points at which ap-

[17]Results are qualitatively similar if we use the full sample to estimate the baseline specification.

[18]These L1 normed distances are weighted by the fraction of observations in the sample observed at each observed experience profile.

[19]See Appendix C.2.1.

proximate cross-partial derivatives can be estimated. For example, if the combinations (2,1,1,1), (2,2,1,1), (2,1,2,1), and (2,2,2,1) are all observed somewhere in the sample, then we can use the smoothed profile above to estimate the cross-partial derivative $\frac{\partial d(2,1,1,1)}{\partial exp^j \partial exp^l}$ via the difference-in-differences $(\hat{d}(2,2,2,1) - \hat{d}(2,2,1,1)) - (\hat{d}(2,1,2,1) - \hat{d}(2,1,1,1))$.

For each pair of two dimensions, we find all the initial four-dimensional cells at which a cross-partial derivative for the chosen two dimensions can be calculated (the cell $(2,1,1,1)$ above would be an element of the set associated with subject- and level-specific experience), and take a weighted average of the difference-in-difference estimates associated with each experience combinations in the set.

For example, let $\mathcal{D}$ represent the set of experience cells observed in the sample. Then

$$\begin{aligned}\mathcal{P}^{j,l} = \{(exp^t, exp^j, exp^l, exp^{sl}) : (exp^t, exp^j, exp^l, exp^{sl}) &\in \mathcal{D}, (exp^t, exp^j + 1, exp^l, exp^{sl}) \\ &\in \mathcal{D}, (exp^t, exp^j, exp^l + 1, exp^{sl}) \\ &\in \mathcal{D}, (exp^t, exp^j + 1, exp^l + 1, exp^{sl}) \in \mathcal{D}\}\end{aligned} \tag{21}$$

represents the set of cells at which a cross-partial derivative between subject- and level-specific experience may be calculated.

We estimate

$$\begin{aligned}\overline{\frac{d(*)}{\partial exp^j, \partial exp^l}} = \sum_{k \in \mathcal{P}^{j,l}} \omega_k [(&\hat{d}(exp^t_k, exp^j_k + 1, exp^l_k + 1, exp^{sl}_k) - \hat{d}(exp^t_k, exp^j_k + 1, exp^l_k, exp^{sl}_k)) \\ &- (\hat{d}(exp^t_k, exp^j_k, exp^l_k + 1, exp^{sl}_k) - \hat{d}(exp^t_k, exp^j_k, exp^l_k, exp^{sl}_k))].\end{aligned} \tag{22}$$

The weight ω_k is composed of the product of four sub-weights associated with each of the cells included in the difference-in-difference estimate. Each sub-weight represents the fraction of all teacher-school-subject-level-year cells that featured the chosen experience combination. The ω_k

are then re-scaled to sum to 1.

Averaging these derivative estimates across all levels of the $d(*)$ function obscures the possibility that two dimensions may be complements at some experience cell levels and substitutes at others. However, such averaging is necessary to obtain a sufficient degree of precision, and provides a useful general sense of the magnitude of submodularity or supermodularity. Table 11 provides the matrix of these cross-partial averages for all two-dimension pairs. The table reveals fairly strong sub-modularity among general experience, subject-specific experience, and level-specific experience. On average, increasing level experience by one year reduces the return to an additional year of general (subject) experience by $.036$ ($.031$) student-level standard deviations.

Such strong sub-modularity has the potential to explain the negative subject-level experience profile observed in Table 5. High values of both subject-specific and level-specific experience will necessarily be strongly correlated with high values of subject-level-specific experience. If subject-specific experience is the least valuable when paired with high values of level-specific experience, then OLS will best fit the additively separable baseline specification to the observed data by setting the subject-specific profile to capture the high returns to subject-specific experience that exist at lower values of level-specific experience, and then choosing negative values of subject-level specific experience to offset the overstatement of teacher productivity that will occur when both subject- and level-specific experience are high.

To test this potential explanation, we construct a table comparable to Table 5 that instead utilizes the smoothed non-parametric experience production function. For each initial value of each component of experience, we calculate the average marginal effect of an extra year of the chosen component experience (holding the other experience components fixed). The average is taken over all combinations of the other three experience dimensions that are observed in an experience cell that features the chosen initial value in the chosen dimension.

For example, let $\mathcal{Q}^{j,v} = \{(exp^t, exp^j, exp^l, exp^{sl}) : exp^j = v, (exp^t, exp^j, exp^l, exp^{sl}) \in \mathcal{D}, (exp^t, exp^j+$

$1, exp^l, exp^{sl}) \in \mathcal{D}\}$ represent the set of experience cells at which a partial derivative for subject-specific experience at initial value v may be calculated. Then the average marginal effect of subject-specific experience from initial value v can be calculated via:

$$(23)\quad \overline{\frac{\partial d(exp^t, v, exp^l, exp^{sl})}{\partial exp^j}} = \sum_{k \in \mathcal{Q}^{j,v}} w_k[\hat{d}(exp^t_k, v+1, exp^l_k, exp^{sl}_k) - \hat{d}(exp^t_k, v, exp^l_k, exp^{sl}_k)]$$

The weight w_k is composed of the product of two sub-weights associated with the two cells included in the partial derivative estimate. Each sub-weight represents the fraction of all teacher-school-subject-level-year cells that featured the chosen experience combination. The w_k are then re-scaled to sum to 1.

These marginal effects are then accumulated into a profile along each experience dimension, so that the sum of the marginal effects accumulated as of year t can be compared to a teacher with 0 experience in the chosen dimension (as in Table 5).

The results of this exercise are displayed in Table 12. The large negative effects of subject-level experience turn slightly positive once additive separability is not imposed. Furthermore, modest but non-negligible gains to level-specific experience appear as well. A teacher with 5-6 years of level-specific experience is now estimated to be .05 student-level standard deviations more effective than he/she would be with no level-specific experience and identical values of the other three dimensions of experience.

However, we still find that the bulk of the gains from experience stem from general and subject-specific experience, so that the basic qualitative conclusions drawn from the baseline results remain largely unchanged. Indeed, the gains from both subject and general experience are even larger under this methodology than in Table 5. Note, though, that the sum across columns of the year t row of Table 12 no longer captures the predicted productivity as of year t of a teacher who has taught the same subject-level every year, since the cross-partial derivatives, previously restricted to be 0 everywhere, were shown above to be quite negative.

7 Gauging the Magnitude of Achievement Gains from Efficient Use of Context-Specific Teacher Experience

7.1 Methodology

The sizeable gains to subject-specific experience suggest that rotating teachers across subjects could potentially result in non-trivial efficiency losses. To gauge the magnitude of such losses, in this section we present a counterfactual simulation in which we project the performance gains that could be achieved statewide if each principal exploited the full value of the accumulated stock of context-specific experience of the members of his or her teaching staff. To ensure that the simulation captures feasible reallocations, we hold fixed the number of classrooms of each subject-level combination at the levels that actually prevailed at each school in each year. Furthermore, we also hold fixed the total number of classrooms taught by each teacher in each year, since principals may have been constrained in the workload they could assign to their more experienced teachers.[20]. Also, because we do not observe the full teaching histories of any teacher who began teaching before the sample begins in 1995, we do not reallocate the classrooms taught by such teachers. Thus, the efficiency gains produced by our simulation will be a lower bound on the true efficiency gains available to be reaped (though this lower bound will increase toward the true predicted efficiency gain as we move through years of the sample).

Note, however, that although we only observe student test scores in the 10 tested subjects, we

[20]For example, these teachers may also have been teaching untested classes, or performing other valuable services to the school, such as lunchroom monitoring, advising student clubs, or coaching student athletic teams)

observe the full set of subject assignments for each teacher. Hence, we can thus construct post-1995 teaching histories across all standard subjects in North Carolina (such as English 2, or Calculus). Consequently, we can accurately update general and level experience stocks for each teacher, even in years in which they do not teach the tested subjects. However, the results only contains efficiency gains from reallocating classrooms in which the tested subjects were taught, since these are the subjects on which the general and context-specific experience profiles were estimated. We also do not reallocate classrooms in which English 1 was taught, since our field-specific results above suggested minimal gain to context-specific experience in English classes.

To see how such a counterfactual would be implemented, consider the allocation of teachers to classrooms that takes place at a particular school over the set of years in our sample. In theory, we might want to solve the dynamic problem of choosing sequences of yearly allocations to maximize the average test score performance over the entire sample. However, a principal source of dynamic gains would stem from principals experimenting to learn each teacher's comparative teaching advantages. The results presented in the previous section, though, suggest that subject-specific and level-specific permanent teaching talent is quite small relative to either context-specific experience or permanent teaching skill that is general across contexts. Consequently, to highlight the potential gains from exploiting context-specific experience, we ignore any potential efficiency gains from matching teachers to their permanent comparative advantages in the simulations below[21].

Furthermore, solving the dynamic problem requires specifying principal's expectations about the probability that each teacher will remain at the school in each future year as well as expectations about the number of classrooms they will need to fill in each subject-level in each future year. This is particularly problematic for the last few years of the sample, where we cannot observe what will happen.

Consequently, we instead simulate the dynamic effects of re-solving the static optimization problem each year in which the expected average test score is maximized in each year, taking the set

[21] specifically, we impose $\mu_{srjl} = \mu_r \forall srj \in \mathcal{SRJ}$

of classrooms and teachers to be matched in each year as exogenously given at the start of the year. While this necessarily understates the true gains to dynamic optimization, it represents an allocation rule that principals can easily implement each fall without making any projections about enrollment and teacher attrition.

To formulate the static problem, let $\mathcal{J}$ represent the set of subjects, $\mathcal{L}$ represent the set of levels, and let $\mathcal{JL}$ be the set of subject-level combinations. Let C_{jl} be the number of classes to be staffed in subject-level combination $jl \in \mathcal{JL}$, with $N_c = \sum_{jl \in \mathcal{JL}} C_{jl}$ denoting the total number of classes to be staffed. Let $\mathcal{R}$ represent the set of teachers, with R elements. As before, ex_r^{tot} captures the number of years in which teacher r has taught any classroom, ex_r^j captures the number of years in which teacher r has taught at least one classroom in subject j, ex_r^l captures teacher r's years of experience teaching level l, and ex_r^{jl} captures teacher r's experience teaching the subject-level combination jl. Student contributions $X_{it}\beta$ can be ignored, since they are assumed to be constant across counterfactual reallocations (and are assumed to be additively separable from teacher inputs).

Using the estimated smoothed non-parametric experience production function introduced in Section 8, we can predict a counterfactual performance of teacher r in course c at time $t = 0$ via:

$$\hat{\overline{Y}}^c_{r0} = \hat{d}(exp^{tot}_{r0}, exp^{j(c)}_{r0}), exp^{l(c)}_{r0}, exp^{jl(c)}_{r0}) \tag{24}$$

The goal is to choose the mapping $f : \mathcal{C} \rightarrow \mathcal{R}$ from classrooms to teachers that maximizes the sum of student test scores, subject to the constraints that each teacher can only teach in as many classrooms as they were observed teaching in at time t (denoted $\overline{C}_r$), and every classroom must be taught by exactly one teacher:

$$\begin{aligned} &\max_{f:\mathcal{C}\rightarrow\mathcal{R}} \sum_{c\in\mathcal{C}} \hat{\overline{Y}}^{c}_{f(c)} \\ &s.t. \ \sum_{r} 1(f(c)=r) = 1 \ \forall \ c \\ &s.t. \ \sum_{c} 1(f(c)=r) = \overline{C}_r \ \forall \ r \end{aligned} \tag{25}$$

where $1(f(c) = r)$ indicates that teacher r taught course c.

This optimization problem can be recast as a binary integer programming problem:

$$\begin{aligned} &\max_{\mathbf{x}} \mathbf{a} * \mathbf{x} \\ &s.t. \ M_c * \mathbf{x} = 1 \ \forall \ c \\ &s.t. \ N_r * \mathbf{x} = \overline{C}_r \ \forall \ r \\ &s.t. \ \mathbf{x} \in \{0, 1\} \end{aligned} \tag{26}$$

$\mathbf{a}$ consists of a $1x(C * R)$ row vector of predicted student performances for each potential teacher-classroom combination:

$$\mathbf{a} = \begin{pmatrix} \hat{\overline{Y}}^{1}_{1} & \dots & \hat{\overline{Y}}^{1}_{C} & \hat{\overline{Y}}^{2}_{1} & \dots & \hat{\overline{Y}}^{2}_{C} & \dots & \hat{\overline{Y}}^{R}_{1} & \dots & \hat{\overline{Y}}^{R}_{C} \end{pmatrix}$$

$\mathbf{x}$ consists of a $(C * R)x1$ vector of potential teacher assignments:

$$\mathbf{x} = \begin{pmatrix} x_1^1 \\ \vdots \\ x_C^1 \\ x_1^2 \\ \vdots \\ x_C^2 \\ \vdots \\ x_1^R \\ \vdots \\ x_C^R \end{pmatrix}$$

where $x_c^r = 1(f(c) = r)$, an indicator for whether teacher r is assigned to classroom c.

M_c consists of a $1xC * R$ row vector capturing the number of teachers assigned to classroom c (restricted to be 1 $\forall$ c):

$$M_c = \left(\underbrace{\overbrace{0 \ldots 0}^{c-1} \ 1 \ \overbrace{0 \ldots 0}^{C-c}}_{\text{repeated R times}} \ldots \overbrace{0 \ldots 0}^{c-1} \ 1 \ \overbrace{0 \ldots 0}^{C-c} \right)$$

N_r consists of a $1xC * R$ row vector capturing the number of classrooms taught by teacher r (restricted to be equal to $\overline{C}_r$, the number taught in the sample):

$$N_r = \left(\overbrace{0 \ldots 0}^{(r-1)*C} \underbrace{1 \ldots 1}_{C} \overbrace{0 \ldots 0}^{(R-r)*C} \right)$$

We solve this binary integer programming problem for each school in the first year of the sample.

We then update each teacher's context-specific experience profile for the second year given the experience they gained under the optimal assignment in the first year.[22] We repeat this process until the end of the sample so as to reap the long-run rewards associated with accumulating high levels of relevant context-specific experience.

While this procedure captures the gains that could have been reaped by the end of each year had the principal maximized the value of context-specific experience in each school starting in 1995 (the first year of the sample), note that the small payoff in the first few years conflates the fact that past switching has limited potential gains from re-optimizing with the fact that relatively few teachers are being reallocated (because we do not observe the classroom assignment histories for the vast majority of the teachers in the first few years). Thus, we focus on efficiency gains among classrooms assigned in the last 5 years of the sample, when a substantial fraction of teachers are eligible for reassignment

We also compare the results of the dynamic simulation to a static simulation that solves the binary integer programming problem in each year t holding fixed observed teacher assignments up through $t-1$. These results reflect the payoff to the first year of optimal reallocation. The static simulation serves to illustrate the decomposition of gains into the part stemming from initial reassignment to better match teachers' context-specific experience to the classrooms they teach and the part stemming from longer run gains associated with the specialization of the teacher work force.

7.2 Results from Counterfactual Simulations

Table 13 presents the fraction of classrooms whose teacher assignments in the simulation differed from the actual teacher assigned in the data, for both the static and dynamic simulations. Because the scope for efficiency gains from matching and specialization increases in the size of the

[22]Since non-tested subjects are not reallocated, any general or level-specific experience teachers accumulated in those subjects under the true allocation is also included in the update.

teaching force, the classroom reallocation rate is presented separately by number of teachers in the school-field-year combination eligible to be reallocated (i.e. the number who taught at least one classroom in that school-field-year combination in the actual data for whom the full teaching history is observed).[23]. The fraction of classrooms reassigned in the simulations is fairly stable across math, science, and social studies classes. The fraction reallocated rises nearly monotonically in the number of teachers eligible for reallocation. For the static simulation, efficient allocation requires 20 percent of classrooms to be reassigned in two-teacher fields, with the fraction rising to around 45 percent for fields with seven or more teachers. The reallocation rate is only slightly higher in the dynamic simulation, rising from around 25 percent for two-teacher fields to over 50 percent for fields with seven or more teachers.

The high rate of reallocation even in the static simulations suggests that there is a considerable amount of excess mobility among teachers across levels and subjects beyond what is necessary to staff the courses offered, given the set of teachers available. However, the simulations may be reallocating a large number of classes to achieve a negligible efficiency gain. Thus, Table 14 shows the student-weighted average expected test score gain from optimal reallocation among all school-field-year combinations, where school-field-year combinations are grouped, as above, by the number of teachers available to be reallocated.

The efficiency gains for two-teacher fields are between .005 and .009 test score standard deviations in the dynamic simulation. Small gains are not surprising for this case, given the limited scope for specialization and the relatively low reallocation rates observed in Table 14. However, the gains grow fairly rapidly in the number of teachers being reallocated. Four-teacher fields reap efficiency gains of .012 to .018 standard deviations from optimal reallocation, while fields with seven or more teachers reap gains between .013 and .025, with the largest gains typically occurring in social studies.

[23]In the case where only one teacher is certified to teach all of the courses in the field, there can be no gains from teacher reallocation. Thus, school-field-years featuring only one teacher are omitted from Table 13

On one hand, these magnitudes are clearly not large enough to dramatically shift the distribution of student acheivement; even a .025 test score gain is only enough to move an average student from the 50th to the 51st percentile of the state test score distribution. However, a number of other considerations suggest a more optimistic interpretation of these efficiency gains.

First, note that these gains are virtually costless: no change in existing staff is required, and no teacher's load is being changed. It is rare to find across-the-board gains from policy changes that require so little upheaval. Indeed, given that the vast majority of the test-score variation is within classes, most other school-level policies are likely to have similarly sized impact. For example, consider a policy that aims to identify and replace the worst 10 percent of teachers with new hires. Using the estimates from Table 8, a teacher at the 10th percentile of general skill reduces test scores by about -.17 test score standard deviations, so that if such teachers teach only 10 percent of students, average test scores would increase by .017 standard deviations even under the optimistic assumption that replacement teachers were of average quality.

Second, note that the vast majority of students are taught in high schools feature 7 or more teachers in a field. The small numbers of teachers eligible for reallocation in many of the schools in our simulations were driven by two factors. First, we required that full teacher assignment histories were observed (information that could be easily ascertained for all teachers by a principal in an actual school). Second, we only reallocated classrooms in tested courses, so that, for example, teachers who only taught Calculus were not available for reallocation. Thus, the largest efficiency gains from our simulations are probably the relevant gains in most situations, and in fact may still be underestimates for most large schools.

Third, our dynamic simulated gains may be further understated because we keep teacher-school matches fixed as they were in the data. If the simulated allocations had actually been realized, the teacher transfer and hire patterns would have likely evolved in a way that preserves more of the efficiency gains from specialization. [24]

[24]For example, a teacher who split time between Biology and Chemistry is instead allocated only to Biology class-

Of course, our simulated efficiency gains could overstate the true gains if, for example, teachers have a taste for variety, and quit more frequently if they are forced to teach the same subject-level combination repeatedly. Similarly, while our results suggest a small role for permanent context-specific skill, it is possible that we are reallocating teachers away from their permanent comparative advantages.[25]

8 Conclusions

This paper introduces and implements a method for decomposing worker productivity into task-specific and general components of both experience and persistent talent. For high school teachers, the bulk of productivity gains from experience are specific to the tasks (subjects) that the teacher has taught, while the bulk of permanent talent is general across all subject-level combinations.

Since the variation in general talent and the value of subject-specific experience are similar in magnitude, effective personnel management for high school administrators requires a mix of selecting/deselecting of teachers combined with retention of teachers who have considerable experience in a particular subject. Since neither level-specific skill nor level-specific experience seems to be important for teacher productivity, honors classes may be used as a non-pecuniary reward for effective teaching or other undesirable tasks (lunch duty?) without any efficiency loss, to the extent that teachers prefer to teach them. Thus, in addition to the practical importance of learning how to better manage the existing stock of public school teachers, the teacher context also represents a case in which allowing the task-specificity of worker productivity to vary across permanent and experience components turns out to be critical for correctly determining the optimal recruitment

rooms in the simulation, but then moves to a school to replace a retiring Chemistry teacher. Chemistry students at the new school will be predicted to perform worse under the simulation than under the original allocation. However, if our assignment algorithm had actually been employed by schools, a different teacher would have been hired to teach Chemistry.

[25] If these comparative advantages are known ex ante, one should be able to ensure that teachers become specialists in the subjects or levels in which they already had a comparative advantage.

and assignment policy for an organization's workers.

Note, though, that the results of the decomposition we estimate may not generalize to other occupations or even to alternative definitions of teachers' tasks. In particular, the set of tasks we considered were still fairly similar in scope. For example, we might observe greater variation in task-specific talent among teachers if we included productivity as a high school athletic coach as one of a teacher's tasks. Similarly, developing students' cognitive and non-cognitive skills might represent two different tasks facing a teacher even within a given classroom context, and teachers good at teaching abstract concepts may not be good at handling student emotional crises. Fortunately, a similar decomposition may be estimated in any context in which worker productivity may be measured at the task level, and where the mix of tasks changes over time.

Appendices

A Identification of Experience Profiles

To see how identification might be secured, a simple case in which there are only two subjects, chemistry (C) and physics (P), and only two difficulty levels, basic (B) and honors (H). Suppose that four different teachers (not necessarily at the same school) each teach a different subject-level combinations in their first year: Teacher 1 teaches basic physics (BP) in her first year, while teacher 2 teaches honors physics (HP), teacher 3 teaches basic chemistry (BC) and teacher 4 teaches honors chemistry (HC). Suppose then that all four teach honors chemistry (HC) every year thereafter. To keep the example simple, suppose further that each of components of experience is fully persistent (no depreciation), and that each teacher only teaches classes in one subject-level per year. Figure 1 displays the course assignment paths taken by each teacher, along with the observed stocks of general, course-specific, level-specific, and subject-level specific experience that teachers will possess at the beginning of each of their school years.

Consider a difference-in-difference estimator that compares the change in teacher 1's average student biology test scores between years 2 and 3 with the corresponding change for teacher 2. Since both teachers are teaching the same subject-level (HC) in each of the two years, focusing on changes differences out the permanent general and context-specific skills of the two teachers. Furthermore, comparing across teachers removes the common gains from the second year of (previous) general experience and the first year of subject-specific and subject-level specific experience. Because teacher 2 taught at the honors level in her first year, the extent to which teacher 1's performance converges to or diverges from teacher 2's performance between years 2 and 3 will reflect the relative value of the 2nd year of level-specific experience compared to the 1st year: $(d^l(2) - d^l(1)) - (d^l(1) - 0)$.[26] If instead we compare the change in student performance between

[26]Note that since returns to experience can only be identified relative to other levels of experience, we must normal-

years 3 and 4 for the same two teachers (1 and 2), we recover the relative value of the 3rd year of level-specific experience compared to the 2nd year: $(d^l(3)-d^l(2))-(d^l(2)-d^l(1))$. Indeed, conditional on knowing the value of the first year of experience, $d^l(1)$, we can trace out the entire path of returns to level-specific experience by comparing the divergence/convergence in the performance of teachers 1 and 2 as they progress through their careers. If we replacing teacher 2 with teacher 3 in the comparisons above, we instead trace out the path of returns to subject-specific experience. Now that the returns to subject-specific and level-specific experience have been identified, replacing teacher 3 with teacher 4 identifies the path of returns to subject-level-specific experience. Finally, the growth path of teacher 4, who never switched subjects or levels, identifies the path of returns to general experience (conditional on knowing $d^t(1)$).

To see how the value of the first year of each component of experience might be identified, consider a second scenario in which teacher 1 teaches the following sequence of courses in her first four years: $BC \rightarrow HC \rightarrow BP \rightarrow HC$. Teacher 2 teaches the same set of courses, but in a different sequence: $BP \rightarrow HC \rightarrow BC \rightarrow HC$. Figure 2 illustrates the stocks of general and context-specific experience each teacher possesses during each year of teaching. Since both teachers teach honors chemistry with the same accumulated experience profile in year 4, comparing the performance of the two teachers identifies the difference in permanent teaching skill between the two teachers (part of which may be honors-chemistry specific): $\mu^2_{CH} - \mu^2_{CH}$. Once relative permanent skill has been identified, comparing the same two teachers' average student residuals in year 2 (when both were teaching honors-chemistry) identifies the return to the 1st year of subject-specific experience, $d^j(1)$. Replacing basic chemistry with honors physics in this example would instead identify the return to the 1st year of level-specific experience ($d^l(1)$), while replacing it with honors chemistry would identify the return to the 1st year of subject-level specific experience($d^{jl}(1)$). The return to the first year of general experience ($d^t(1)$) can then be identified via the growth in student average residuals from the 1st to the 2nd year from teachers who teach the same subject-level in each of their first two years.

ize one value for each function. We do so by setting $d^k(0) = 0$ for $k \in \{tot, j, l, lk\}$.

Note that the experience profiles in the examples above were all identified by comparing teachers' performance growth across years in which the same subject-level combination was taught. Because the average performance of each teacher in each school-subject-level combination is perfectly fit by the unrestricted school-teacher-subject-level and school-subject-level fixed effects, these cell averages provide no identifying variation for the experience profiles. Thus, the inclusion of these fixed effects forces the identification of the experience profiles to be delivered exclusively from the path of productivity growth within school-teacher-subject-level combinations.

Also, while the sample histories used in the scenarios above are stylized, note that there are many alternative moments that also provide identifying variation. Indeed, given the frequency with which subject and level switching occurs, we frequently observe multiple teachers who have taught the same set of subjects and levels over their careers at the school, but have taught them in different orders, or in different proportions. Since each different sequence also implies a different pattern of potential depreciation for a given model of depreciation, such comparisons allow us in principle to simultaneously estimate the rates at which different experience components depreciate.[27]

Furthermore, each subject or level switch, regardless of the point in the career, provides a further source of identifying variation for the various context-specific experience profiles. Consequently, not only are these experience profiles estimatable with reasonable precision (at least for the first several years of experience), but there are myriad overidentifying tests that can be implemented if one worries that particular sequences may be likely to occur in conjunction with particular changes in unobserved inputs (in violation of Assumption 1). Indeed, in Section 8 we show that the function linking 4-dimensional stocks of general and context-specific teacher experience to student performance is non-parametrically identified, and we present estimates from a more flexible (though noisily estimated) specification.

[27]In practice, after some experimentation, we include in our estimated specifications 4 dummy variables indicating whether the teacher taught the current subject last year, the current level last year, the current subject-level last year, and whether the teacher taught any class last year.

B Identification of Permanent Teaching Skill

To illustrate how $\hat{\mu}_{srjl}$ can be identified given any of Assumptions 2A-2C paired with 3A-3C, consider the a teacher r' who teaches subject j' and level l' in school s' during year t. Suppose, without loss of generality, that the teacher has a vector of general and context specific experience of $(ex^{tot} = 4, ex^{j} = 3, ex^{l} = 2, ex^{jl} = 2)$. The expected residual performance of her students in subject-level combination (j, l), denoted $Z_{ict} = Y_{ict} - X_{ict}\beta$, is given by:

$$\begin{aligned} & E[Z_{ict}|(s,r,j,l) = (s',r',j',l'), (ex^{tot}_{rt}, ex^{j}_{rt}, ex^{l}_{rt}, ex^{jl}_{rt}) = (4,3,2,2)] \\ & = \delta_{sjl} + \mu_{srjl} + d^{tot}(4) + d^{j}(3) + d^{l}(2) + d^{jl}(2) \end{aligned} \tag{27}$$

Since the experience profiles $d^{tot}(*)$, $d^{j}(*)$, $d^{l}(*)$, and $d^{jl}(*)$ were identified using variation in the sequence of course assignments in the subsection above, moments that capture the level of performance at a point in time identify $\delta_{sjl} + \mu_{srjl}$. Under Assumption 2B, $\delta_{sjl} = 0 \; \forall \; (s, j, l)$, so such moments identify μ_{srjl} directly. Under Assumption 2A, we can use the fact that the (student weighted) average teacher quality in each school-subject-level is assumed to be zero. Specifically, the average residual performance of students in a particular school-subject-level is given by:

$$E[Z_{ict}|(s,r,j,l) = (s',r',j',l')] = \delta_{sjl} + \overline{d^{tot}(ex^{tot})} + \overline{d^{j}(ex^{j})} + \overline{d^{l}(ex^{l})} + \overline{d^{jl}(ex^{jl})}, \tag{28}$$

which identifies δ_{sjl}. To identify δ_{s} under Assumption 2C, we simply average at the school level instead of the school-subject-level level. Thus, μ_{srjl} can be identified for each combination of school-teacher-subject-level that we actually observe in the data.

C Robustness Check Technicalities

C.1 Testing for Endogenous Responses to School-Year Shocks

To test for endogenous responses to school-year-shocks, recall that if we use residuals in which the estimated experience profiles have not been removed, then school-year average residuals will naturally be correlated with the experience profile composition of the teachers in the school-year via the causal effect of teacher context-specific experience.

This problem can be solved, however, by re-weighting the classroom residual averages that compose school-year averages to account for differences in experience profile composition across school-years. To see how this might be done, let Z_{ict} represent student i's residual test score in classroom c at time t, where the predicted effect of all inputs in the baseline model except teacher experience have been removed:

$$\begin{aligned} Z_{ict} &= Y_{ict} - X_{ict}\hat{\beta}_{jl} - \hat{\delta}_{sjl} - \hat{\mu}_{srjl} \\ &= \hat{d}^{tot}(exp_{rt}^{tot}) + \hat{d}^{j}(exp_{rt}^{j}) + \hat{d}^{l}(exp_{rt}^{l}) + \hat{d}^{jl}(exp_{rt}^{jl}) + \hat{\epsilon}_{ict} \end{aligned} \tag{29}$$

We can form student-weighted school-year average residuals by weighting the average residuals of the classrooms in the school-year by the number of students they contained:

$$\begin{aligned} \overline{Z}_{st} &= \frac{1}{C_{st}} \sum_{c\in(s,t)} w_c \overline{Z}_c \\ &= \frac{1}{C_{st}} \sum_{c\in(s,t)} w_c \overline{\hat{d}}_c^{tot} + \overline{\hat{d}}_c^{j} + \overline{\hat{d}}_c^{l} + \hat{d}_c^{jl} + \overline{\hat{\epsilon}}_c \\ &\approx \frac{1}{C_{st}} \sum_{c\in(s,t)} w_c \overline{\hat{d}}_c^{tot} + \overline{\hat{d}}_c^{j} + \overline{\hat{d}}_c^{l} + \hat{d}_c^{jl} + \phi_{st} \end{aligned} \tag{30}$$

where $w_c = \frac{N_c}{N_{sy}}$, and we have assumed for simplicity that classroom, teacher-year, and student error components average to approximately zero within a school-year.[28]

Equation 30 makes clear that schools which have a disproportionate fraction of classrooms taught by teachers with low stock of context-specific experience will tend to have negative school-year average residuals. However, we can remove this effect of experience composition by replacing w_c with $w_{exp(c)}$, where $w_{exp(c)}$ is the fraction of classrooms in the full sample featuring teachers with the same experience profile as the teacher who taught classroom c. Imagine for now that each school-year featured the full support of teaching profiles (though perhaps with different frequencies). Then the reweighted school-year average residual yields:

$$\tilde{Z}_{st} = \sum_{c \in (s,t)} w_{exp(c)} \overline{Z}_c \approx \sum_{exp \in \mathcal{EX}} w_{exp} \overline{\hat{d}}^{tot}(exp) + \overline{\hat{d}}^{j}(exp) + \overline{\hat{d}}^{l}(exp) + \hat{d}^{jl}(exp) + \phi_{st}$$

$$= K + \phi_{st} \tag{31}$$

where K is a constant that reflects the average contribution of teacher experience in the full sample (given the normalization chosen). Thus, the reweighted school-year average residual gives us an unbiased estimator of the school-year deviation in quality from the school's long run average. Given the set $\{\tilde{Z}_{st}\}$, we can then examine whether particular experience levels disproportionately occur during schools' relatively ineffective years by taking a profile-specific weighted average of $\tilde{Z}_{st}$, where the weight on each school-year is the fraction of classrooms in the full sample featuring the experience profile in question that appeared within that school-year:

$$E[\phi_{st} | exp^k_{rt} = exp'] = \sum_{st \in \mathcal{ST}} w_{st}(exp')(\tilde{Z}_{st} - K) \tag{32}$$

[28]To the extent that there are relatively few teacher-years some school-years, our test statistic also will reflect endogenous responses to teacher-year shocks, another potential threat to validity.

Unfortunately, each four-dimensional experience profile is not observed in each school-year (a failure of common support), so that we cannot fully purge the effect of experience profile composition within a school-year by reweighting observed classroom averages. We approximate the true reweighted average as best by distributing the weight that would have been placed on unobserved profiles to observed profiles based on the L1 distance between the unobserved and observed profiles (passed through a normal kernel to smooth this distribution).

While this method increases substantially the weight placed on profiles in the underrepresented region of the distribution of four-dimensional experience profiles, in school-years where all of the teachers have relatively low predicted experience component of productivity, no amount of reweighting will possibly allow the observed experience composition of the school-year to approximate the full sample distribution. However, this failure of common support biases our test against us, since our test statistic will identify spurious differences in average school-year shocks across profiles. Thus, our reweighting estimator allows us to place an upper bound on the bias produced from endogenous school-year.

C.2 Misspecification Tests: Testing the Additive Separability of Context-Specific Experience Profiles

C.2.1 Addressing Sampling Error

To address this issue, we assume that the true structural function $d(*,*,*,*)$ is differentiable everywhere, and smooth our estimates using a normal density with zero mean and standard deviation 0.5 as a kernel, via:

$$\tilde{d}(\mathbf{ex}) = \frac{\sum_{\mathbf{ex}' \in \mathcal{EX}} w_{ex'} \phi(|\mathbf{ex}' - \mathbf{ex}|)\hat{d}(\mathbf{ex}')}{\sum_{\mathbf{ex}' \in \mathcal{EX}} w_{ex'} \phi(|\mathbf{ex}' - \mathbf{ex}|)} \tag{33}$$

The magnitude $|\mathbf{ex}' - \mathbf{ex}|$ that serves as the argument to the normal density is the L1 norm or taxicab distance between the two experience profiles: $|\mathbf{ex}' - \mathbf{ex}| = |ex^{tot'} - ex^{tot}| + |ex^{j'} - ex^{j}| + |ex^{l'} - ex^{l}| + |ex^{jl'} - ex^{jl}|$. The weight $w_{ex'}$ represents the fraction of observations in the sample in which the experience profile ex' is observed. Thus, the impact that $\hat{d}(0,0,0,0)$ has on $\tilde{d}(0,0,0,1)$ will be greater than that of $\hat{d}(0,1,0,1)$, despite equal L1 distances, because $\hat{d}(0,0,0,0)$ is much more precisely estimated than $\hat{d}(0,1,0,1)$. The chosen bandwidth yields a four-dimensional function $\tilde{d}(*,*,*,*)$ that is smooth enough to remove considerable sampling error, yet is still flexible enough to reveal true complementarities where they may occur.

References

Aaronson, Daniel, Lisa Barrow, and William Sander (2007) 'Teachers and Student Achievement in Chicago Public Schools.' *Journal of Labor Economics* 25(1), 95–135

Atteberry, Allison, Susanna Loeb, and James Wyckoff (2013) 'Do first impressions matter? improvement in early career teacher effectiveness.' Technical Report, National Bureau of Economic Research

Becker, Gary Stanley (1964) *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education, by Gary S. Becker,...* (London)

Boyd, Donald, Hamilton Lankford, Susanna Loeb, Jonah Rockoff, and James Wyckoff (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

Clotfelter, Charles, Helen Ladd, and Jacob Vigdor (2007) 'Teacher Credentials and Student Achievement: Longitudinal Analysis with Student Fixed Effects.' *Economics of Education Review* 26(6), 673–682

Clotfelter, Charles T, Helen F Ladd, and Jacob L Vigdor (2006) 'Teacher-student matching and the assessment of teacher effectiveness.' *Journal of Human Resources* 41(4), 778–820

Görlich, Dennis, and Andries De Grip (2009) 'Human capital depreciation during hometime.' *Oxford Economic Papers* 61(suppl 1), i98–i121

Hanushek, Eric A, and Steven G Rivkin (2010) 'Generalizations about using value-added measures of teacher quality.' *The American Economic Review* 100(2), 267–271

Hanushek, Eric, John Kain, Daniel O'Brien, and Steven Rivkin (2005) 'The Market for Teacher Quality.' NBER Working Paper 11154, National Bureau of Economic Research, Inc

Harris, Douglas, and Tim R. Sass (2006) 'Value-Added Models and the Measurement of Teacher Quality.' Unpublished Manuscript

Harris, Douglas N (2009) 'Would accountability based on teacher value added be smart policy? an examination of the statistical properties and policy alternatives.' *Education* 4(4), 319–350

Harris, Douglas N, and Tim R Sass (2011) 'Teacher training, teacher quality and student achievement.' *Journal of public economics* 95(7), 798–812

Jackson, C. Kirabo (2012) 'Do High School Teachers Really Matter?' NBER Working Paper 17722, National Bureau of Economic Research, Inc

_ (2013) 'Match Quality, Worker Productivity, and Worker Mobility: Direct Evidence from Teachers.' *Review of Economics and Statistics*

Jackson, Clement Kirabo, and Elias Bruegmann (2009) 'Teaching students and teaching each other: The importance of peer learning for teachers.' *American Economic Journal: Applied Economics* 1(4), 85–108

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

Mansfield, Richard (2013) 'Teacher Quality and Student Inequality.' Working Paper

Mincer, Jacob, and Haim Ofek (1982) 'Interrupted work careers: Depreciation and restoration of human capital.' *Journal of Human Resources* pp. 3–24

Ost, Ben (2013) 'How Do Teachers Improve? The Relative Importance of Specific and General Human Capital.' *American Economic Journal - Applied Economics (Forthcoming)*

Papay, John P, and Matthew A Kraft (2011) 'Productivity returns to experience in the teacher labor market: methodological challenges and new evidence on long-term career growth.' Technical Report, Working Paper

Rivkin, Steven, Eric Hanushek, and John Kain (2005) 'Teachers, Schools, and Academic Achievement.' *Econometrica* 73(2), 417–458

Rockoff, Jonah (2004) 'The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data.' *American Economic Review: Papers and Proceedings of the One Hundred Sixteenth Annual Meeting of the American Economic Association* 94(2), 247–252

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

Sass, Tim, Jane Hannaway, Zeyu Xu, David Figlio, and Li Feng (2010) 'Value Added of Teachers in High-Poverty Schools and Lower Poverty Schools: Implications for Research, Policy and Management.' Working Paper 52, National Center for Analysis of Longitudinal Data in Education Research

Wiswall, Matthew (2013) 'The Dynamics of Teacher Quality.' *Journal of Public Economics* 100, 61–78

9 Tables and Figures

Table 1: Experience Distribution among Classes taught by 2nd and 3rd year Teachers

	Years of Experience			Fraction of
	Course	Level	Crs-Lvl	Classes
Second-Year Teachers	1	1	1	25.0%
	1	1	0	28.6%
	1	0	0	20.6%
	0	1	0	19.1%
	0	0	0	6.7%
	Total classes taught by a 2nd-year teacher:			**280,140**
Third-Year Teachers	2	2	2	17.8%
	2	2	1	2.1%
	2	2	0	15.5%
	2	1	1	2.6%
	2	1	0	10.8%
	2	0	0	10.7%
	1	2	1	5.4%
	1	2	0	8.9%
	1	1	1	3.0%
	1	1	0	5.2%
	1	0	0	2.5%
	0	2	0	9.9%
	0	1	0	2.7%
	0	0	0	2.9%
	Total classes taught by a 3rd-year teacher:			**244,965**

Notes: The table presents classrooms in our final estimation sample having teachers in either their second or third year teaching. Notice that since multiple classes can be taught in a year, this explains seemingly impossible experience combinations such as the second row for second-year teachers. For instance, a teacher who taught both remedial chemistry and honors biology in year one would have this type of experience profile teaching remedial biology in year two.

Table 2: Teacher Mobility Across Courses: Regression Sample

	Course										
Course	Algebra 1	Algebra 2	Biology	Chemistry	Civics	E/L/P	English	Geometry	Physics	Physical Sciences	U.S. History
Algebra 1	**1,780**	594	29	8	5	9	20	637	15	36	17
	1	0.334	0.016	0.004	0.003	0.005	0.011	0.358	0.008	0.020	0.010
Algebra 2	594	**741**	6	3	2	3	3	380	10	8	4
	0.802	1	0.008	0.004	0.003	0.004	0.004	0.513	0.013	0.011	0.005
Biology	29	6	**754**	133	7	20	25	7	41	339	19
	0.038	0.008	1	0.176	0.009	0.027	0.033	0.009	0.054	0.450	0.025
Chemistry	8	3	133	**369**	1	1	2	4	82	213	1
	0.022	0.008	0.360	1	0.003	0.003	0.005	0.011	0.222	0.577	0.003
Civics	5	2	7	1	**440**	141	9	2	0	5	245
	0.011	0.005	0.016	0.002	1	0.320	0.020	0.005	0.000	0.011	0.557
E/L/P	9	3	20	1	141	**572**	32	3	0	22	311
	0.016	0.005	0.035	0.002	0.247	1	0.056	0.005	0.000	0.038	0.544
English	20	3	25	2	9	32	**892**	7	0	16	27
	0.022	0.003	0.028	0.002	0.010	0.036	1	0.008	0.000	0.018	0.030
Geometry	637	380	7	4	2	3	7	**855**	9	10	5
	0.745	0.444	0.008	0.005	0.002	0.004	0.008	1	0.011	0.012	0.006
Physics	15	10	41	82	0	0	0	9	**169**	114	1
	0.089	0.059	0.243	0.485	0.000	0.000	0.000	0.053	1	0.675	0.006
Physical	36	8	339	213	5	22	16	10	114	**734**	16
Sciences	0.049	0.011	0.462	0.290	0.007	0.030	0.022	0.014	0.155	1	0.022
U.S.	17	4	19	1	245	311	27	5	1	16	**783**
History	0.022	0.005	0.024	0.001	0.313	0.397	0.034	0.006	0.001	0.020	1

Notes: The top entry in the (i,j)-th cell is the number of teachers who are observed teaching in both the i-th and the j-th difficulty level (not necessarily in the same year). The bottom entry of the (i,j)-th cell is the fraction of teachers ever observed teaching the i-th difficulty level who are also observed teaching the j-th difficulty level at some point during the sample.

Table 3: The Pattern of Teacher Mobility Across Difficulty Levels

		Difficulty Level	
		Low	High
Diff. Level	Low	**5,956**	2,974
		1	0.499
	High	2,974	**3,003**
		0.990	1

Notes: The top entry in the (i,j)-th cell is the number of teachers who are observed teaching in both the i-th and the j-th difficulty level (not necessarily in the same year). The bottom entry of the (i,j)-th cell is the fraction of teachers ever observed teaching the i-th difficulty level who are also observed teaching the j-th difficulty level at some point during the sample.

Table 4: The Pattern of Teacher Mobility Across Course-Levels for the Mathematics Field

Course-Level		Algebra 1 Low	Algebra 1 High	Algebra 2 Low	Algebra 2 High	Geometry Low	Geometry High
Algebra 1	Low Level	**1,767**	71	551	238	587	248
		1	0.040	0.312	0.135	0.332	0.140
	High Level	71	**84**	20	15	23	17
		0.845	1	0.238	0.179	0.274	0.202
Algebra 2	Low Level	551	20	**682**	236	330	138
		0.808	0.029	1	0.346	0.484	0.202
	High Level	238	15	236	**295**	128	78
		0.807	0.051	0.800	1	0.434	0.264
Geometry	Low Level	587	23	330	128	**786**	270
		0.747	0.029	0.420	0.163	1	0.344
	High Level	248	17	138	78	270	**339**
		0.732	0.050	0.407	0.230	0.796	1

Notes: The top entry in the (i,j)-th cell is the number of teachers who are observed teaching in both the i-th and the j-th difficulty level (not necessarily in the same year). The bottom entry of the (i,j)-th cell is the fraction of teachers ever observed teaching the i-th difficulty level who are also observed teaching the j-th difficulty level at some point during the sample.

Table 5: The Effect of General, Course, Level, and Course-Level Experience on Student Test Scores

Years Experience	General	Subject	Level	Subj.-Level	Combined
	(1)	(2)	(3)	(4)	(5)
1 yr	0.056***	0.023***	-0.009	0.005	0.076***
	[0.011]	[0.010]	[0.010]	[0.009]	[0.003]
2 yrs	0.071***	0.047***	-0.009	-0.006	0.103***
	[0.015]	[0.014]	[0.014]	[0.013]	[0.004]
3 yrs	0.072***	0.070***	-0.010	-0.023*	0.109***
	[0.019]	[0.018]	[0.018]	[0.018]	[0.005]
4 yrs	0.077***	0.076***	-0.016	-0.027*	0.111***
	[0.022]	[0.022]	[0.022]	[0.021]	[0.005]
5-6 yrs	0.081***	0.077***	-0.013	-0.038*	0.108***
	[0.025]	[0.025]	[0.026]	[0.025]	[0.006]
7-10 yrs	0.088***	0.080***	-0.027	-0.050**	0.091***
	[0.030]	[0.030]	[0.031]	[0.030]	[0.007]
11-14 yrs	0.063**	0.092***	0.010	-0.088**	0.077***
	[0.036]	[0.038]	[0.037]	[0.039]	[0.013]

$N = 747,890$ student-class observations. Robust standard errors are in brackets. Significance at the 10%, 5%, and 1% levels are represented by ***, **, and * respectively.

Table 6: The Effect of Total and Course Experience on Student Test Scores (Restricted Specification)

Years Experience	Restricted Spec.		Standard Spec.
	Total	Subject	Total Only
	(1)	(2)	(3)
1 yr	0.048***	0.027***	0.073***
	[0.006]	[0.005]	[0.0034]
2 yrs	0.063***	0.041***	0.102***
	[0.008]	[0.008]	[0.0039]
3 yrs	0.061***	0.050***	0.109***
	[0.010]	[0.009]	[0.0044]
4 yrs	0.063***	0.049***	0.114***
	[0.012]	[0.011]	[0.0049]
5-6 yrs	0.068***	0.040**	0.115***
	[0.014]	[0.013]	[0.0051]
7-10 yrs	0.064***	0.029**	0.104***
	[0.017]	[0.016]	[0.0059]
11-14 yrs	0.060***	0.020	0.093***
	[0.021]	[0.022]	[0.0088]

$N = 747,890$ student-class observations. Robust standard errors are in brackets. Significance at the 10%, 5%, and 1% levels are represented by ***, **, and * respectively.

Table 7: Subject-Specific Experience Profiles by Field

Years	Math		Science		Social Studies		English	
Exp.	Total	Subj.	Total	Subj.	Total	Subj.	Total	Subj.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1 yr	0.062***	0.021***	0.062***	0.039***	0.055***	0.029***	0.025*	0.004
	[0.011]	[0.009]	[0.011]	[0.011]	[0.012]	[0.011]	[0.016]	[0.016]
2 yrs	0.079***	0.035***	0.078***	0.053***	0.085***	0.041***	0.022	0.021
	[0.014]	[0.012]	[0.015]	[0.014]	[0.017]	[0.015]	[0.023]	[0.022]
3 yrs	0.083***	0.056***	0.081***	0.056***	0.081***	0.050***	0.027	0.008
	[0.017]	[0.015]	[0.019]	[0.018]	[0.020]	[0.019]	[0.028]	[0.027]
4 yrs	0.083***	0.053***	0.061***	0.066***	0.109***	0.048**	0.036	0.002
	[0.019]	[0.018]	[0.022]	[0.021]	[0.024]	[0.022]	[0.032]	[0.032]
5-6 yrs	0.109***	0.043**	0.061***	0.062***	0.108***	0.042*	0.020	-0.009
	[0.022]	[0.021]	[0.025]	[0.025]	[0.028]	[0.027]	[0.037]	[0.037]
7-10 yrs	0.125***	0.032	0.029	0.061**	0.101***	0.036	0.016	-0.016
	[0.027]	[0.026]	[0.031]	[0.031]	[0.035]	[0.033]	[0.044]	[0.044]

$N = 747,890$ student-class observations. Robust standard errors are in brackets. Significance at the 10%, 5%, and 1% levels are represented by ***, **, and * respectively.

Table 8: True Variances in Fixed Effects

	Lower Bound (1)		Intermediate (2)		Upper Bound (3)	
	Var.	SD	Var.	SD	Var.	SD
School-Subject-Level-Teacher FE	0.0145	0.120	0.0321	0.179	0.0450	0.212
School-Teacher FE	0.0115	0.107	0.0277	0.166	0.0406	0.201
Subject-Level Deviations	0.0029	0.054	0.0044	0.067	0.0044	0.067
School-Subject-Teacher FE	0.0136	0.117	0.0308	0.175	0.0438	0.209
Level Deviations	0.0009	0.030	0.0012	0.034	0.0012	0.034
School-Level-Teacher FE	0.0126	0.112	0.0294	0.171	0.0423	0.206
Subject Deviations	0.0019	0.044	0.0027	0.052	0.0027	0.052
Subject-Level Residual	0.0001	0.014	0.0004	0.021	0.0004	0.021

Estimates derived from the estimated variance less the student weighted average of standard errors in our fixed effect estimates.

Table 9: Estimates of the Bias in Experience Profiles Due to Endogenous Teacher Assignment Responses to School-Year Shocks

Years Experience	Total	Course	Level	Course-Level
	(1)	(2)	(3)	(4)
1 yr	0.002	0.002	-0.006	0.003
	[0.003]	[0.003]	[0.003]	[0.003]
2 yrs	0.001	0.007	-0.004	0.002
	[0.004]	[0.004]	[0.004]	[0.004]
3 yrs	-0.005	0.006	-0.007	0.001
	[0.005]	[0.004]	[0.005]	[0.004]
4 yrs	-0.010	0.012	0.006	-0.006
	[0.005]	[0.005]	[0.005]	[0.005]
5-6 yrs	-0.003	0.008	0.002	-0.002
	[0.005]	[0.006]	[0.005]	[0.006]
7-10 yrs	-0.009	0.006	0.003	-0.001
	[0.006]	[0.007]	[0.006]	[0.007]

$N = 747,890$ student-class observations. Robust standard errors are in brackets. Significance at the 1%, 5%, and 10% levels are represented by ***, **, and * respectively.

Table 10: Testing Dynamic Teacher Assignment

	Total		**Subject**		**Level**		**Subj-Lvl**	
	Next Year	Ever	Next Year	Ever	Next Year	Ever	Next Year	Ever
1-2 Yrs	-0.043	-0.041	-0.048	-0.052	-0.057	-0.081**	-0.052*	-0.076***
	[0.048]	[0.044]	[0.036]	[0.033]	[0.040]	[0.037]	[0.030]	[0.028]
	N=2,273		N=3,342		N=2,934		N=4,217	
1-3 Yrs	0.091**	0.046	0.009	-0.028	-0.021	0.013	0.016	0.011
	[0.040]	[0.057]	[0.028]	[0.044]	[0.032]	[0.047]	[0.027]	[0.038]
	N=1,327		N=1,838		N=1,695		N=2,226	
1-4 Yrs	0.117**	-0.122*	0.003	0.079	-0.025	0.044	-0.018	0.104**
	[0.054]	[0.070]	[0.042]	[0.057]	[0.043]	[0.063]	[0.041]	[0.049]
	N=833		N=1,009		N=1,039		N=1,200	
1-5 Yrs	0.132**	0.074	0.112*	0.119	0.048	-0.089	-0.027	0.028
	[0.066]	[0.099]	[0.059]	[0.078]	[0.057]	[0.081]	[0.051]	[0.065]
	N=598		N=650		N=733		N=770	
1-6 Yrs	-0.146*	0.034	0.037	0.230**	0.073	0.151	-0.070	-0.029
	[0.088]	[0.124]	[0.082]	[0.099]	[0.073]	[0.108]	[0.081]	[0.085]
	N=417		N=407		N=489		N=459	
2 Yr Pooled	-0.031	-0.033	-0.019	-0.022	-0.042*	-0.061***	-0.018	-0.040**
	[0.027]	[0.024]	[0.021]	[0.019]	[0.023]	[0.021]	[0.018]	[0.016]
	N=8,425		N=10,372		N=10,262		N=12,108	
4 Yr Pooled	0.063**	-0.033	-0.011	0.019	-0.003	0.033	-0.022	0.043
	[0.025]	[0.038]	[0.024]	[0.032]	[0.022]	[0.033]	[0.023]	[0.027]
	N=3,638		N=3,682		N=4,239		N=4,007	

Table 11: Average Cross-Partial Effects Derived from Non-Parametric Experience Production Function

	Total (1)	Subject (2)	Level (3)	Subj.-Level (4)
Total	–	-0.019	-0.036**	0.009
	–	[0.015]	[0.017]	[0.018]
Subject	-0.019	–	-0.031***	-0.012
	[0.015]	–	[0.013]	[0.010
Level	-0.036**	-0.031***	–	0.003
	[0.017]	[0.013]	–	[0.012]
Subj.-Level	0.009	-0.012	0.003	–
	[0.018]	[0.010]	[0.012]	–

$N = 747,890$ student-class observations. Standard errors are in brackets, and were computed using the delta method. Significance at the 1%, 5%, and 10% levels are represented by ***, **, and * respectively.

Table 12: Average Accumulated Marginal Effects Derived from Non-Parametric Experience Production Function

Years Experience	Total	Subject	Level	Subj.-Level
	(1)	(2)	(3)	(4)
1 yr	0.016***	0.029***	0.032***	0.014***
	[0.004]	[0.012]	[0.008]	[0.004]
2 yrs	0.065***	0.056***	0.036***	0.021***
	[0.024]	[0.019]	[0.012]	[0.008]
3 yrs	0.090**	0.089***	0.032**	0.021**
	[0.040]	[0.026]	[0.016]	[0.011]
4 yrs	0.086**	0.111***	0.042**	0.018*
	[0.044]	[0.033]	[0.019]	[0.014]
5-6 yrs	0.117**	0.115***	0.049***	0.016
	[0.062]	[0.043]	[0.021]	[0.016]
7-10 yrs	0.110*	0.123***	0.034*	0.016
	[0.071]	[0.048]	[0.023]	[0.018]
11-14 yrs	0.109	0.127**	0.048**	0.004
	[0.100]	[0.058]	[0.026]	[0.022]

$N = 747,890$ student-class observations. Standard errors are in brackets, and were computed using the delta method. Significance at the 1%, 5%, and 10% levels are represented by ***, **, and * respectively.

Table 13: Counterfactual Simulations: Fraction of Classrooms Reallocated

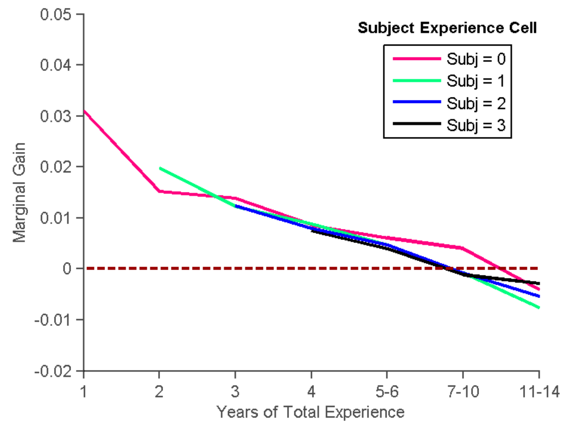
	Math		Science		Social Studies	
# Eligible Teachers	**Static** (1)	**Dynamic** (2)	**Static** (3)	**Dynamic** (4)	**Static** (5)	**Dynamic** (6)
2 tch	0.214	0.238	0.175	0.282	0.192	0.268
3 tch	0.294	0.335	0.275	0.385	0.323	0.393
4 tch	0.385	0.432	0.320	0.454	0.378	0.442
5-6 tch	0.383	0.439	0.364	0.479	0.424	0.485
7-10 tch	0.421	0.450	0.414	0.475	0.484	0.554
11+ tch	0.437	0.513	—	—	0.489	0.535

Table 14: Counterfactual Simulations: Achievement Gains

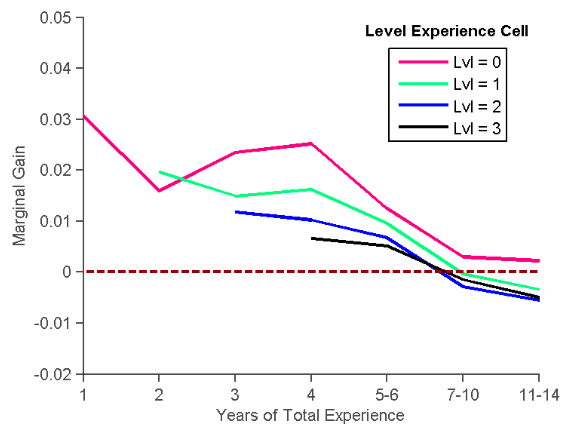
	Math		**Science**		**Social Studies**	
# Eligible Teachers	**Static** (1)	**Dynamic** (2)	**Static** (3)	**Dynamic** (4)	**Static** (5)	**Dynamic** (6)
1 tch	—	-0.001	—	0.002	—	0.002
2 tch	0.005	0.006	0.004	0.009	0.005	0.009
3 tch	0.007	0.007	0.008	0.013	0.009	0.015
4 tch	0.011	0.012	0.008	0.013	0.012	0.018
5-6 tch	0.012	0.011	0.010	0.015	0.015	0.021
7-10 tch	0.014	0.013	0.009	0.017	0.019	0.025
11+ tch	0.018	0.016	—	—	0.025	0.023

Figure 1: A Graphical Depiction of Complementarity Between Dimensions of Experience (Part 1)

(a) Gains to General Experience Across Values of Subject-Specific Experience



(b) Gains to General Experience Across Values of Level-Specific Experience



(c) Gains to Subject-Level Experience Across Values of General Experience

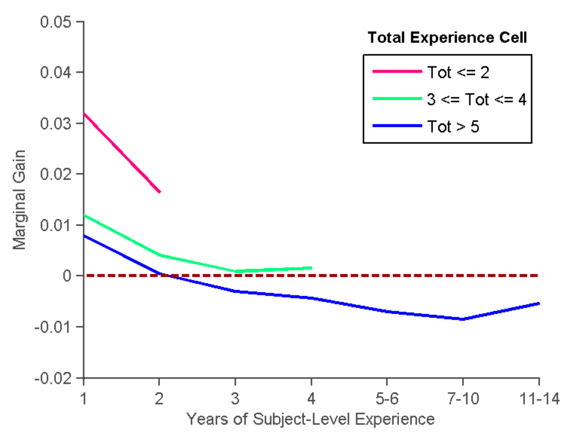
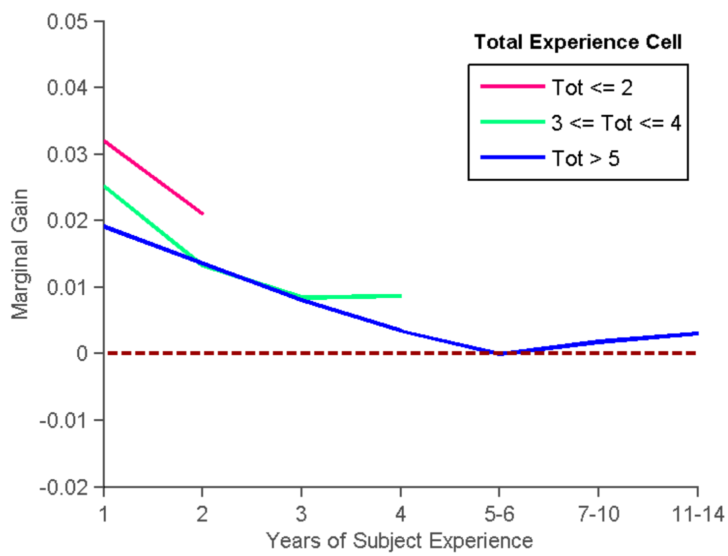


Figure 2: A Graphical Depiction of Complementarity Between Dimensions of Experience (Part 2)

(a) Gains to Subject Experience Across Values of General Experience



(b) Gains to Subject Experience Across Values of Level-Specific Experience

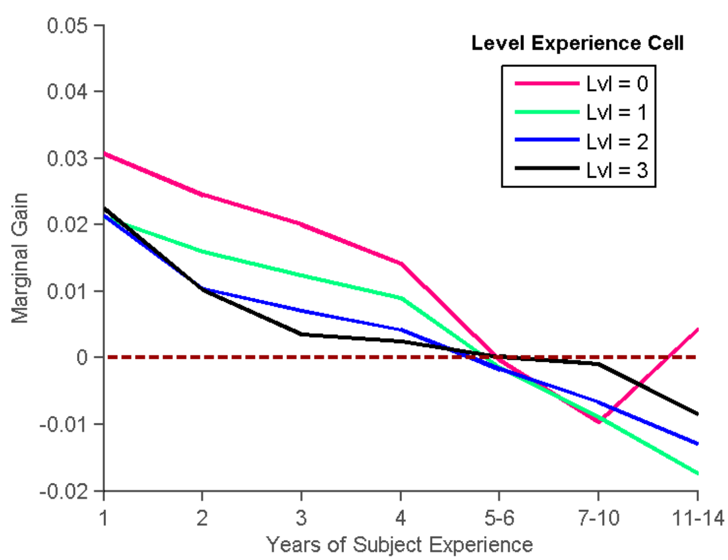
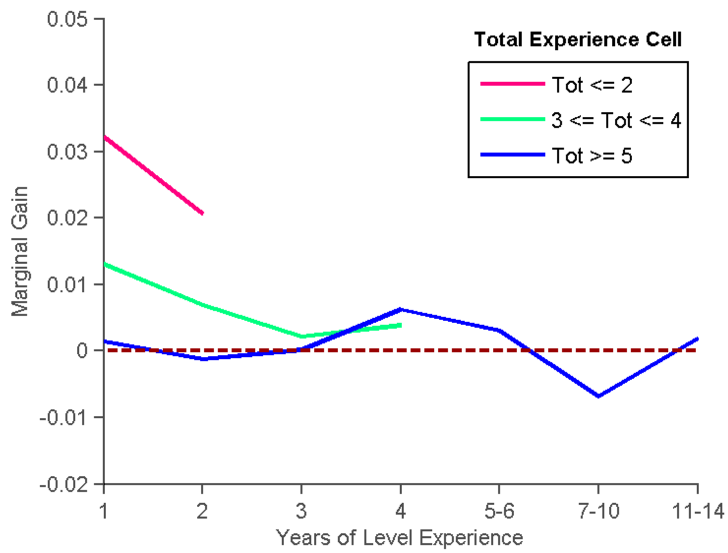


Figure 3: A Graphical Depiction of Complementarity Between Dimensions of Experience (Part 3)

(a) Gains to Level Experience Across Values of General Experience



(b) Gains to Level Experience Across Values of Subject Experience