The Determinants of Teachers' Occupational Choice*

PRELIMINARY: Cite at your own risk.

Kevin Lang† and Maria Dolores Palacios‡

March 19, 2018

Abstract

Among college graduates, teachers have both low average AFQT and high average risk aversion. Using a dynamic optimization model with unobserved heterogeneity, we find that the low mean AFQT score among teachers primarily reflects a low return to other skills, correlated with AFQT, rather than a low return to cognitive skill within teaching. The compression of earnings within teaching attracts relatively risk-averse individuals. Were it possible to make teacher compensation mimic the return to skills and riskiness of the non-teaching sector, overall compensation in teaching would increase. Moreover, such a shift would substantially reduce the utility of many current teachers, making the process of reform challenging. Importantly, our conclusions are sensitive to the degree of heterogeneity for which we allow, and even a model with no unobserved heterogeneity appears to fit well within sample. It would be easy to conclude that allowing for two or three types fits the data adequately. Formal methods reject this conclusion. The BIC favors seven types. Ranking models using cross-validation, eight types is better although the improvements from going from six to seven, from seven to eight types and from eight to nine types are noticeably smaller than those from adding an additional type to a lower base. Importantly, the results of policy exercises are very sensitive to the degree of heterogeneity included in the model.

*We are grateful to Peter Arcidiacono, Richard Blundell, Ivan Fernandez-Val, Hiro Kaido, Zhongjun Qu and Marc Rysman for helpful discussons and to participants in in the Cowles Foundation Conference on Structural Microeconomics and seminars at Boston University, Brandeis University, Carleton University, the Federal Reserve Bank of Minneapolis; the International Symposium on Contemporary Labor Economics (Jinan University); Queen's University, the University of Connecticut, the University of Western Ontario and Washington University in Saint Louis for their comments and suggestions. Lang acknowledges NSF funding under grant SES-1260917. The usual caveat applies.

†Boston University, NBER and IZA. email: lang@bu.edu

‡Boston University. email: doloresp@bu.edu.

1 Introduction

We make two contributions. On the substantive side, we examine the feasibility of a policy that makes the earnings structure for teachers more closely resemble that outside of teaching. On the methods side, we show that standard tests of in-sample fit can lead us to accept models with inadequate allowance for heterogeneity.

Teacher salaries are typically based on salary scales that depend only on education, experience and seniority and not on measures of quality or achievement. Yet we know that, at least in other settings (e.g. Lazear (2000)), tying compensation more closely to performance can both increase the productivity of a fixed set of individuals and attract more productive individuals. Hoxby and Leigh (2004) provide support for this hypothesis in teaching (see also Bacolod (2007)).[1] Consequently, there is considerable interest in performance pay for teachers (see for example, the National Research Council report, Hout and Elliott (2011)).

At the same time, teachers unions have typically resisted performance pay. This is not entirely surprising. As we will show, teachers are, on average, more risk averse than the general population of college graduates. Standard theory implies that they will therefore require greater compensation to offset the increased risk associated with performance pay.

In this paper, we do not examine performance pay, *per se*, but we ask how the composition of the teaching profession would change if education and ability were compensated in the same way as in the general labor market for college graduates, presumably making teaching a similarly risky occupation. Like much of this literature, we look at the effect on general ability as measured by test scores and/or potential earnings outside teaching since we do not have a measure of teacher effectiveness. We take it as given that ability within and outside teaching are correlated, albeit imperfectly. This is supported by our estimates of a strong correlation between potential earnings in teaching and non-teaching.

To do this we estimate a dynamic model of occupation choice in which individuals decide each year whether to work as a teacher, in some other occupation, or not to be employed. The decision is dynamic because there are sector-specific returns to experience and because

[1] Leigh (2012) finds evidence that higher pay increases the test scores of students in teacher training programs in Australia and some evidence that greater earnings dispersion outside of teaching lowers scores.

there is a cost of moving among the sectors.[2]

Our model is closest in both format and spirit to Stinebrickner (2001b) and particularly Stinebrickner (2001a). There are, however, some important differences in our treatment of uncertainty and of variation in the importance of earnings in the utility function. In addition, unlike him, we do not limit our sample to individuals who obtain their teaching qualification early on but consider all college graduates. The policy changes of interest to us may affect the decision to obtain a teaching qualification. We are also able to follow individuals much later into their careers which allows us to consider exit from and reentry into teaching.

Having measured and unmeasured ability rewarded within teaching as they are outside teaching leads to a shift of the more skilled types to teaching and slightly raises the average AFQT of teachers. Our results are consistent with Hanushek et al. (2004) finding only modest effects of earnings on quality.

Even this conclusion ignores the difficulty of effecting a transition. Reform requires transitioning to a compensation system that rewards characteristics differently from the current system and increases risk for a population that is risk averse relative to other college graduates. The reform we study would make a substantial proportion of experienced teachers worse off. Reforms are therefore likely to be very disruptive in the short run, regardless of whether they are beneficial in the long run.[3]

Our conclusions turn out to be very sensitive to the extent to which we allow for unobserved individual heterogeneity. In doing so, we face two risks. If we do not allow for sufficient heterogeneity, the model is misspecified. If we allow for excess heterogeneity, although the estimates remain consistent, our counterfactual estimates may suffer from overfitting of the original model. Strikingly even the model with no unobserved heterogeneity appears to fit the data well. One could easily conclude that allowing for two or three unobserved types is adequate. But the simulation results are quite different if we allow for four types and dramatically different if we allow for five or six types. With seven or more types, our results are again quite different from those with five or six. To choose the number of types we use

[2]There is an enormous literature on the decisions to become a teacher and to leave teaching which we will not attempt to review thoroughly. This literature is reviewed in Dolton (2006).

[3]See Biasi (2017) for a study of teacher mobility and exit following Wisconsin's Act 10, which radically overhauled the compensation system for teachers in that state.

both the Bayesian Information Criterion (BIC) and cross-validation using different numbers of unobserved types. The BIC favors using seven types out of the nine models compared, while the cross-validation favors eight or more types.[4]

To compare the predicted simulation effects of each model we use the Hausman formula and calculate the standard errors of differences between the predictions of the different models. If the true model has N types, then the most efficient estimate of the policy effect is based on the model with this number of types. Models with fewer than N types are inconsistent while models with more than N types are consistent but not efficient. Thus, under the null that there are N types we can test for equality of the simulation predictions calculating the variance of the difference as the variance of the prediction using N+1 (or, more generally, N+k, k$>$ 0) types minus the variance of the prediction using N types. The results from these comparisons suggest that we require at least seven types. When we move beyond seven types we observe a large loss of precision and we cannot reject that all the average characteristics predicted are the same as those using the seven-type model. We view this as supporting our reliance on the seven-type model.

2 Data and Some Empirical Regularities

We reverse the usual order of presentation and discuss the data before the model because certain regularities will influence how we develop the model.

We use the National Longitudinal Survey of Youth 1979 (NLSY79). Since the NLSY79 is well-known to labor economists, we skip a general description of the survey. We restrict the sample to college graduates and drop observations for years in which individuals report being self-employed, in the military or working fewer than 35 hours a week. Finally, we drop individuals interviewed fewer than four times. Table 1 shows summary statistics for the 1,071 individuals in our sample, divided into three categories: (1) teachers, (2) non-teachers and (3) not working. Note that an individual might be in all three categories over the course of the panel.

[4]The results for nine types are preliminary and suggest that the cross validation will reject using fewer than nine types.

Table 1: Summary statistics

	Teachers	**Non-teachers**	**Not working**
	(1)	(2)	(3)
Earnings	51.1	76.6	-
in $1,000	(1.3)	(1.4)	-
Risk aversion	0.3008	-0.0389	0.0197
standardized	(0.0178)	(0.0071)	(0.0346)
Schooling	17.4	16.8	16.7
in years	(0.0)	(0.0)	(0.0)
AFQT	-0.3061	0.0367	0.0518
standardized	(0.0205)	(0.0069)	(0.0343)
Individuals	220	1050	371
Observations	2,591	20,464	825

NOTES: Standard errors in parenthesis.
*An individual might be in all three categories.

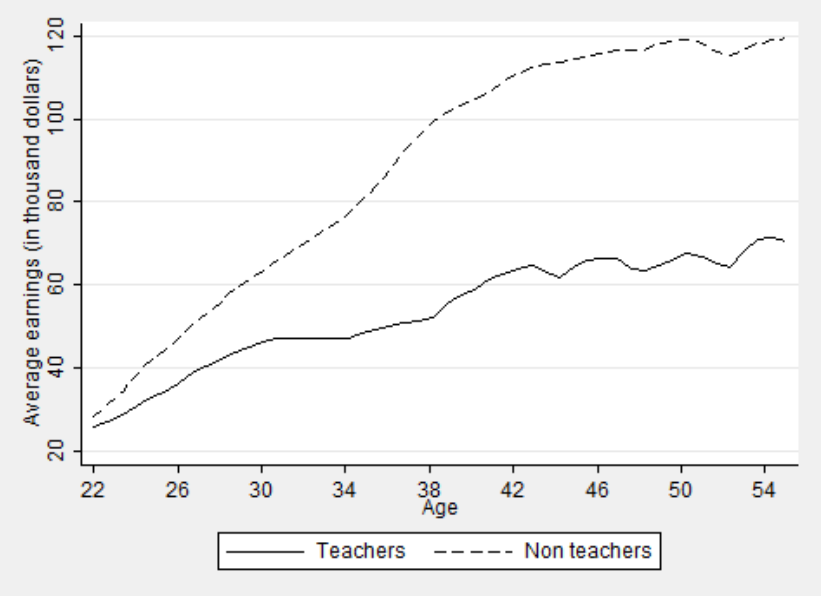
Compensation over the Lifecycle: Table 1 shows that teachers, on average, earn less than other college graduates. As shown in Figure 1a, teachers' and non-teachers' have similar earnings when 22 years old, but a gap emerges as they age. In this and other figures, we show estimates for individuals age 22-55. However, sample sizes at both extremes are small and should therefore be treated with caution. By the time they are 55 years old, non-teachers earn almost 56 thousand dollars[5] more than teachers do. This pattern does not merely reflect the changing composition of the various groups over the lifecycle. Controlling for individual fixed effects or restricting the estimates to individuals who are only teachers or only non-teachers does not substantially change the pattern.

Risk and Uncertainty: The standard deviation of earnings is initially small and similar for teachers and non-teachers, but as age increases, the standard deviation remains modest for teachers and grows dramatically for non-teachers (see Figure 1b). Since the residual could reflect factors known to the individual but not the econometrician, this does not necessarily imply that teaching is less risky than other occupations, but it is suggestive. And, in fact, if we regress log earnings on schooling, experience, year fixed effects and individual fixed

[5]In 2012 dollars.

effects, residual earnings variation is higher and grows faster among non-teachers.

(a)



(b)

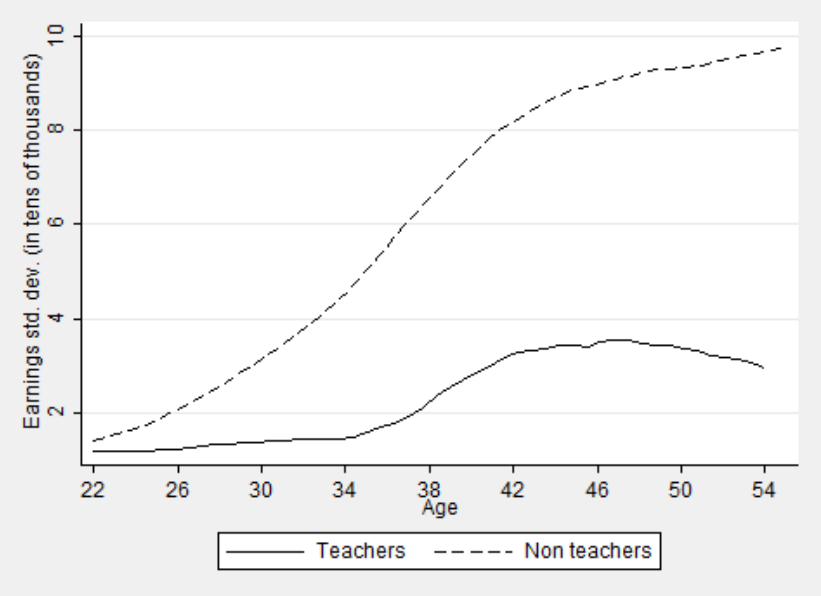


FIGURE 1: Teachers and Non-teachers earnings* by age

(a) Average earnings

(b) Standard deviation of earnings

**Translated into 2012 dollars.*

Risk aversion: Therefore, we would expect risk-averse individuals to sort into teaching.

We construct a risk-aversion parameter using three questions that were asked in each of four years:[6] (1) Would you take a job that could double your income or cut it by 1/2 with a 50-50 chance?, (2) Would you take a job that could double your income or cut it by 1/3 with a 50-50 chance?, and (3) Would you take a job that could double your income or cut it by 1/5 with a 50-50 chance? Using the responses to these questions we construct a risk aversion parameter. We assign a "1" to individuals who responded yes to the three questions, then a "2" to individuals who responded no to question one but yes to the other two, then we assign a "3" to individuals who responded no to questions one and two but yes to the last one, and finally we give a "4" to the most risk averse individuals who responded no to all questions. Because the same questions were asked in several years, the risk aversion parameter for a given individual may change. To have a measure of risk aversion for every year we use the most recent risk aversion parameter available for each individual.

As shown in Figure 2, teachers are more risk averse than individuals working in other occupations.

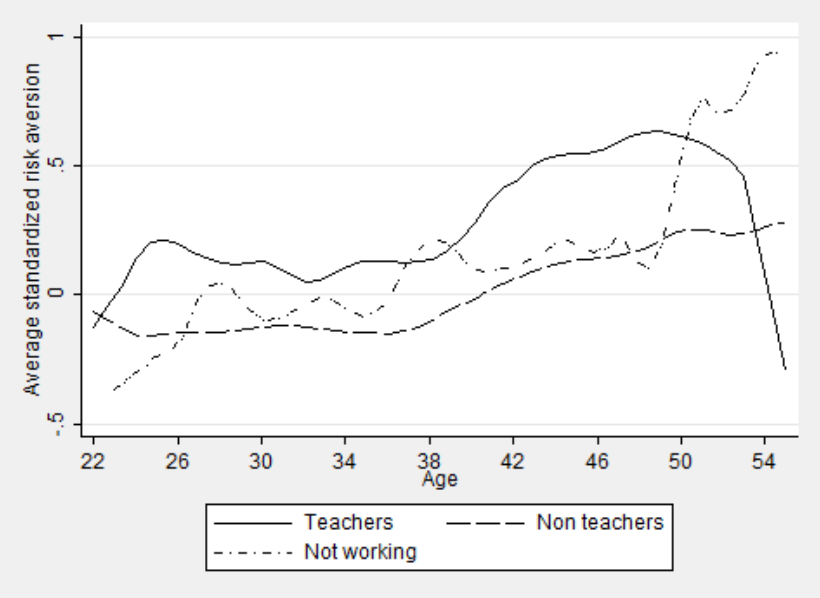


FIGURE 2: Risk aversion by age

Measured skills: Since teachers are rewarded for their education, we would not be surprised if they had higher average levels of education than workers in other occupations

[6] 1993, 2002, 2004 and 2006

even among college graduates. This is confirmed in Table 1. In fact, in 2000 (when our sample is between 35 and 43 years old) 72% of teachers had graduate studies[7] compared with only 43% of non-teachers and 17% of individuals out of the labor force.

However, Figure 3 shows that they generally have less skill as measured by the AFQT. The NLSY79 measured the AFQT when the sample was 15-22 years old. The mean AFQT percentile, adjusted for the age at which the individual took that test, for teachers in our sample is 67 while the mean percentile for non-teachers is 75. Figure 3 shows the standardized average AFQT by age. This "observable ability" measure is lower for teachers of all age groups, perhaps because teachers' earnings are not as responsive to cognitive skills as those for non-teachers. For some age groups this difference is almost 0.5 standard deviations.

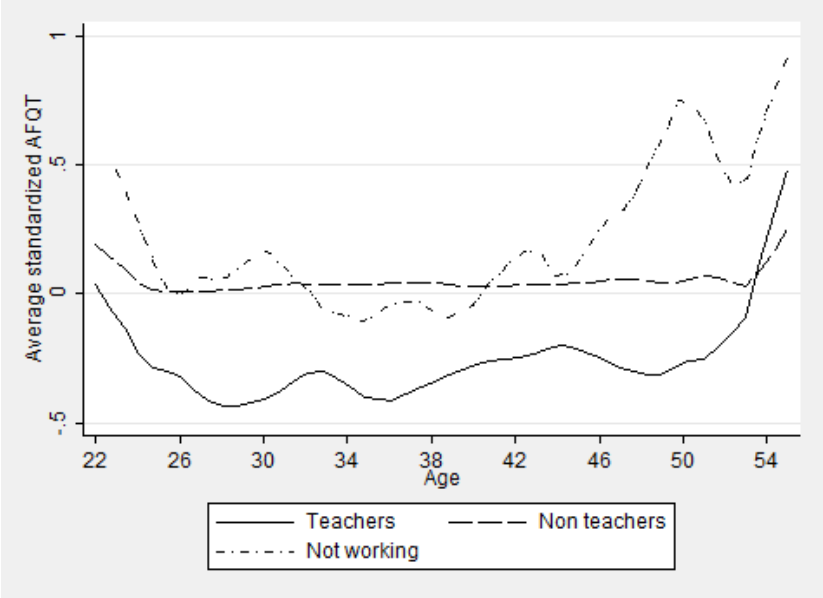


FIGURE 3: AFQT by age

Occupational mobility: Table 2 shows that mobility between teaching and non-teaching is modest. Fully 93.3% of individuals who are teachers in one year continue as teachers the following year, while only 5.2% move to a non-teaching occupation even though non-teachers account for 86% of the person-years in our sample. Mobility from non-teaching to teaching is rarer. As can be seen from the second row of the table, only 0.8% of non-teachers transition to teachers in an average year even though teachers account for 11% of the person-years in

[7]Individuals that have at least 17 years of schooling.

Table 2: Transition matrix (percentages)

		t		
		Teacher	Non-teacher	Not working
$t-1$	Teacher	93.3	5.2	1.5
	Non-teacher	0.8	98.3	0.9
	Not working	4.8	23.5	71.6

our sample.

3 Model and estimation

Utility is quadratic in earnings:

$$u(w) = aw - bw^2 \tag{1}$$

where $a > 2bw$. Letting earnings equal expected earnings plus a mean zero shock with variance σ_ξ^2, expected utility is:

$$E[u] = aE[w] - b(E[w])^2 - b\sigma_\xi^2. \tag{2}$$

Thus, the expected utility of the agent is increasing at a decreasing rate with her expected earnings and decreases with the variance. Although b does not correspond to a standard coefficient of absolute or relative risk aversion, it captures the worker's degree of risk aversion. We remind the reader that b is our data and that the σ_ξ^2 are estimated, up to a factor of proportionality, as parameters.

3.1 Occupation choice

The one-period utility function depends on the decision made, on expected earnings and on the nonpecuniary characteristics of the occupations. We do not model these characteristics since we are modeling the choice of teaching versus non-teaching rather than the choice among admittedly heterogeneous jobs within these broader occupations. If all individuals valued the nonpecuniary components equally, we could capture their value by adding occupation-

specific constants to the utility function. The teaching constant would capture the utility, net of earnings and risk, of working as a teacher relative to not working and similarly for the non-teaching constant. The difference between the coefficients would capture the utility of teaching relative to non-teaching.

Of course, people with different characteristics may value the nonpecuniary characteristics in systematically different ways. To capture these differences, we include additional sociodemographic controls in the utility function. Since we specify different utility functions for teaching and non-teaching except for the effect of earnings, these sociodemographic controls are implicitly interacted with occupation.

Finally, we allow for the possibility that it is costly to transition between states. It may, for example, take time to adjust to the different nature of the working environment in teaching or non-teaching.

Consequently, utility in each state d depends on the state variables $s_{it} = \left\{\bar{Z}_{it}^d, \xi_{it}^d\right\}$ as given by

$$\begin{aligned} U_{it}^d(s_{it}) &= u(d, \bar{Z}_{it}^d) + \xi_{it}^d & (3) \\ &= \theta_{u1} E[w_{it}^d] + \theta_{u2}(E[w_{it}^d])^2 + \theta_{u3}^d z_{it} + \theta_{u4}^d (1 - d_{i,t-1}) + \theta_{u5}^d o_{i,t-1} + \xi_{it}^d & (4) \end{aligned}$$

where $\bar{Z}_{it}^d := \left\{E[w_{it}^d], z_{it}, d_{i,t-1}, o_{i,t-1}\right\}$ represents the observed state variables. $E[w_{it}^d]$ is expected earnings and z_{it} includes individuals' characteristics such as gender, marital status, risk aversion, age, AFQT percentile, schooling[8] and whether the individual has children. $d_{i,t-1}$ is a dummy variable equal to one if individual i chose occupation d in period $t-1$ and zero otherwise. $o_{i,t-1}$ is also a dummy variable, equal to one if individual i chose not to work in period $t-1$ and zero otherwise. Thus, θ_{u4}^d is the cost of transitioning into occupation d from the other occupation and θ_{u5}^d is the cost of transitioning into occupation d from non-employment. We normalize the utility associated with non-employment to be 0. We also set the cost of transitioning to non-employment at 0.

Note that individuals' utility for each occupation differs depending on their risk aversion. One should think of the coefficient on risk aversion as being roughly proportional to σ_ξ^2 in

[8]Schooling is occupation specific. We explain in the details in the next section.

equation (2).[9] Risk aversion is only measured in four surveys. We hold it fixed between these and assume that individuals know their future risk aversion.

The individual's unobserved preferences for each occupation (taste shocks) are given by the ξ_{it}^d. We assume the ξ_{it}^d's are taken from an extreme value type I distribution and temporarily assume that the errors are serially uncorrelated. Later, we will address serial correlation by allowing for multiple types.

We model the occupation choice process as consisting of $(T - \hat{t})$ periods, where T is the retirement date, which we will generally impose to be age 60, and $\hat{t}$ is the first year the individual enters the workforce. Each period, individuals can choose among teaching, other jobs (non-teaching) and non-employment. If they decide to work, their earnings depend on their occupation choice.

Individuals are forward-looking. When choosing an occupation in period t, they take into account not only the one-period utilities associated with the choices but also their effect on sector-specific experience and therefore on future earnings. In addition, they may face a switching cost.

From entry into the workforce until retirement (T), individuals weigh the consequences of their decisions for future utility. A full solution of the dynamic programming problem consists of finding $E[max(V_t^1(s_{it}), V_t^2(s_{it}), V_t^3(s_{it}))]$ at all values of z_{it}, $E(w_{it}^d)$, $d_{i,t-1}$ and $o_{i,t-1}$ for all t, where the choice-specific value function is:

$$V_t^d(s_{it}) = \begin{cases} U_{it}^d + \delta E[V_{t+1}(s_{i,t+1})|s_{it}, d_{it} = d] & \text{if } t < T \\ U_{it}^d & \text{if } t = T \end{cases} \tag{5}$$

The choice-specific value function $V_t^d(s_{it})$ can be decomposed as $v_{it}^d + \xi_{it}^d$, where v_{it}^d is the expected choice-specific value function that has a closed form solution. We set the discount factor δ equal to 0.95.[10] Given the extreme value assumption for the distribution of taste

[9] We say roughly because the measure of risk aversion is related to but not identical to the coefficient b in the expected utility equation.

[10] As commented by Aguirregabiria and Mira (2010) the discount factor in most applications is not estimated because it is poorly identified (e.g., see Rust (1987)).

shocks, the probability of an individual choosing occupation d in period t takes a logit form:

$$P(d_{it} = d|\bar{Z}_{it}, \theta_u) = \frac{exp(v_{it}^d)}{\sum_d exp(v_{it}^d)} \tag{6}$$

where the sums are taken over the three possible options available to the individual. Since we have a finite time horizon, and taste shocks are distributed as an extreme value type I, expected value functions have a closed form analytical expression and can be calculated by backward induction. For a more detailed discussion on estimation of discrete choice dynamic programming models see Aguirregabiria and Mira (2010) and Arcidiacono and Ellickson (2011).

3.2 Earnings

Earnings depend on the occupation chosen and are a function of: a time trend which also captures linear age effects, individual characteristics, and experience in the teaching and non-teaching sectors. Thus log earnings for a given decision $d = \{\text{teacher, non-teacher}\}$ in year t for individual i are given by:

$$\begin{aligned} log(w_{it}^d) &= \theta_w^d \bar{X}_{it}^d + \epsilon_{it}^d & (7) \\ &= \theta_{w1}^d f(trend) + \theta_{w2}^d x_i + \theta_{w3}^d g_{it}^d + \theta_{w4}^d f(expT_{it}) + \theta_{w5}^d f(expN_{it}) + \epsilon_{it}^d & (8) \end{aligned}$$

where x_i includes gender, marital status, race, AFQT score and whether the individual has children.

Schooling, g_{it}, is occupation specific. For any observed decision, schooling is just the number of years of education that individual i has, but for the alternative choices we assume individuals would have at least the same schooling as the contemporaneous average individual in that occupation, that is $g = \max\{actual\, education, average\, education\, in\, occupation\}$. For instance, if individual i decides to be a non-teacher in period t and has sixteen years of schooling then g_{it} for her non-teaching log earnings equation is sixteen. However, if the average schooling of teachers of her age is higher, say eighteen years, then schooling for individual i at time t for her teaching log earnings equation is eighteen. This implies that

individual i would get more schooling if she decided to be a teacher. Since we are not modeling schooling decisions and education is explicitly rewarded in teacher compensation contracts, allowing for education to be higher for off-path decisions is an important feature of a teacher-occupation-choice model.

The occupation-specific experience terms, experience as a teacher ($expT_{it}$) and experience as a non-teacher ($expN_{it}$), evolve depending on the individual's choices. $f(.)$ is a quadratic function. Finally, the shocks (the ϵ_{it}^{d}'s) are unknown to the individual at the time of the decision and are assumed to be normally distributed with mean zero and variance σ_d^2. We want to capture the fact that the variance of earnings increases with age (or experience). This is particularly true for the non-teaching occupation. Therefore, we model the σ_d's as linear functions of age. For identification of the coefficient on earnings in the utility function it is crucial that an exclusion restriction exists, a variable appearing in the earnings equation and only affecting utility through earnings. We use the sector-specific experience terms as the exclusion restriction.

3.3 Heterogeneity, serial correlation and selection

So far, we have assumed that unobserved preferences and unobserved ability are both uncorrelated over time and uncorrelated with each other. Thus an intense unobservable preference for teaching in period t would not be related with having an intense unobservable preference for teaching in period $t+1$. Similarly, there is no persistent unobserved ability which is known to the individual but not to the econometrician.

To address these concerns, suppose that there are L types of people that differ in their preferences for each occupation and in their unobserved abilities.[11] We allow the constant terms of the utility functions and log earnings equations, and the coefficients on the expected earnings terms in the utility functions to vary among types. Thus, the utility and log earnings

[11] See Keane and Wolpin (1997), Eckstein and Wolpin (1999) and Arcidiacono (2004) for other papers that control for unobserved heterogeneity in dynamic discrete choice models. Stinebrickner (2001a) uses this approach in a closely related model of occupational choice by qualified teachers.

equations for type l in occupation d are:

$$\begin{aligned} U^d_{it,l}(s_{it}) &= \omega^{\mathbf{d}}_{\mathbf{u,l}} + \theta_{\mathbf{u1,l}}\mathbf{E}[\mathbf{w^d_{it,l}}] + \theta_{\mathbf{u2,l}}(\mathbf{E}[\mathbf{w^d_{it,l}}])^{\mathbf{2}} + \theta^d_{u3}z_{it} + \theta^d_{u4}(1-d_{i,t-1})... \\ &\quad +\theta^d_{u5}o_{i,t-1} + \xi^d_{it} \end{aligned} \tag{9}$$

$$log(w^d_{it,l}) = \omega^{\mathbf{d}}_{\mathbf{w,l}} + \theta^d_w \bar{X}^d_{it} + \epsilon^d_{it}. \tag{10}$$

Note that, for a given type, we restrict the coefficients on the expected earnings terms in the utility functions to be the same in teaching and non-teaching. Thus expected earnings in teaching and non-teaching occupations give type l the same utility.

3.4 Estimation

We calculate the parameters using maximum likelihood. Without unobserved heterogeneity, the contribution of individual i to the likelihood function is the product of the likelihood contribution of occupation decisions $P(.)$ and the likelihood contribution of earnings $f_w(.)$:

$$L_i(\theta) \equiv \prod_{t=1}^{T_i} P(d_{it}|\bar{Z}_{it},\theta) f_w(log(w_{it})|d_{it},\bar{X}_{it},\theta). \tag{11}$$

To estimate the parameters when we include unobserved heterogeneity we use a mixture distribution, where π_l is the proportion of the lth type in the population. These proportions and the unobservable preferences and abilities are fixed over time, allowing us to control for serial correlation and selection. With unobserved heterogeneity the contribution of individual i to the finite mixture of likelihoods is:

$$l_i(\theta,\Omega,\pi) = log(\sum_{l=1}^{L} L_i(\theta_l,\omega_l)\cdot\pi_l). \tag{12}$$

The set of structural parameters to estimate consists of 64 coefficients when there is no unobservable heterogeneity. For each additional type we include there are seven extra parameters. We estimate the model for one through nine types.

In the remainder of the paper, we primarily present the results without heterogeneity and with seven types since this is the number chosen using the BIC and since the results for

Table 3: Bayesian Information Criterion

	BIC	Difference
One type	24,880	-4,958
Two types	19,922	-1,765
Three types	18,157	-617
Four types	17,540	-142
Five types	17,398	-156
Six types	17,242	-138
Seven types	17,104	1
Eight types	17,014	118
Nine types	17,132	

seven, eight and nine types are similar.

4 Results

4.1 Choosing the number of types

The most obvious approach to model selection, a likelihood ratio (or similar) test, cannot be used because mixture models violate the requisite regularity conditions. Some parameters are not identified under the null. An obvious alternative, Schwarz's Bayesian Information Criterion (BIC),[12] tends to require a very large number of types, perhaps beyond what is numerically feasible in some contexts.

In our case, as shown in section 5, visually our model fits the data well regardless of the number of types. We use two formal approaches to choose among the models. First, we calculate the BIC for each specification (see Table 3). The BIC continues to improves up to the seven types model, but it is worse for the eight and nine types models.

Second, we use a cross-validation approach. Our approach is based on the following logic. A properly specified maximum likelihood model minimizes out of sample prediction error (Hansen and Dumitrescu, 2016). Therefore, if we believe that one of our models is correctly specified, it should be the one that predicts best out of sample. Thus, we randomly divide our sample into two groups, consisting of 80% and 20% of individuals. We re-estimate the

[12] $BIC = -2 \cdot loglikelihood + d \cdot log(N)$, where N is the sample size and d is the total number of parameters.

models using the larger sub-sample. Then, we use the new coefficients and the data from the other 20% of individuals and calculate the log-likelihoods for each of the nine models. We repeat this exercise twenty times and compare the out-of-sample log-likelihoods.

As with the BIC, this cross-validation approach suggests that we require a large number of unobservable types to address heterogeneity adequately. However, when comparing the seven and eight types models the cross-validation approach suggests eight types is better.[13] Still, there is also some evidence that we are approaching the requisite number of types. In all twenty replications, four types does better out-of-sample than three types which in turn fits better than the model with two types which outperforms the model with one type. When we compare seven versus six, the model with more types predicts better out of sample in a clear majority of replications and averaged across replications, but there are replications that give the opposite result. The same is true when comparing six versus five and five versus four. However, when comparing eight and seven types, eight types does better out-of-sample in only 13 of the 20 sub-samples.

Consequently for this draft, we present the results of the model with seven types. For purposes of comparison, we also show the results without unobserved heterogeneity.

4.2 Estimates of the utility function

The results of the selection equations are given in Tables 4 and 5.

Table 4 displays the estimates that are common to all types. Relative to non-employment, being risk averse increases the utility from teaching and reduces the utility from non-teaching. When we control for unobserved heterogeneity with the seven-type model, the point estimates for risk aversion in teachers' and non-teachers' utility functions still suggest that risk averse individuals tend to prefer teaching. In the absence of unobserved heterogeneity, among single individuals, conditional on the other controls, being male strongly increases the preference for nonemployment relative to both teaching and non-teaching with only a small difference between the two types of occupations. Their preference for non-teaching relative to teaching is more pronounced in the model with seven types.

Being married or with children seems to increase the preference for non-employment over

[13] The preliminary results for nine types point out that nine types will probably be preferred to eight types.

teaching, and for teaching over other occupations for both females and males. These coefficients do not change substantially when controlling for unobserved heterogeneity.

Switching occupations and returning to employment is costly. Controlling for unobserved heterogeneity has very little effect on these coefficients. Using the estimates from the model with no unobserved heterogeneity, for the average teacher earning $51,100 switching from non-teacher to teacher is comparable to a decrease of $18,100 in earnings that year and the cost of returning to employment is equivalent to $35,130 in that year's earnings. For the average non-teacher earning $76,600 switching occupation from teacher to non-teacher is comparable to a decrease of $21,800 in earnings and the cost of returning to employment is equivalent to $39,300 in that year's earnings.

Age, age squared, AFQT and schooling were also included in the estimation. The coefficients on AFQT are very similar for teachers and non-teachers. Whereas, more educated individuals prefer to work as a non-teacher.

The panels labeled Type x in Table 5 show the relation between types and utility. Each type except the first has an additional utility it receives relative to the first type from each of the occupations. These panels also show the full utility that each type receives from expected earnings (in 10,000's of 2012 dollars) and its square. With only one type, the marginal utility of earnings is increasing up to almost $200,000, which covers 95% of our observations. When we allow for unobserved heterogeneity, all types except the first and fourth put more weight on earnings at low values than the single type does. The second type only values earnings up to about $118,000 and the sixth only up to about $117,000 while the other types have positive marginal utility of earnings up to $200,000 or higher.

4.3 Estimates of the log earnings equation

Table 6 shows the estimates of the log earnings equations. When we do not consider other forms of heterogeneity, the coefficient on AFQT percentile is lower for teachers. This is still true when we control for unobserved heterogeneity but the difference between the coefficients on AFQT for teachers and non-teachers is smaller in the seven types model than in the one type model.

Table 4: Occupation specific utility function parameters

			One type		Seven types	
			Coefficient	Stand. Error	Coefficient	Stand. Error
All types	Risk aversion	Teachers	0.1042*	(0.0600)	0.0960	(0.0732)
		Non-teachers	-0.0061	(0.0396)	0.0151	(0.0473)
	Male dummy	Teachers	-1.1840***	(0.2853)	-1.9248***	(0.4268)
		Non-teachers	-1.3417***	(0.2075)	-1.1990***	(0.2651)
	Age	Teachers	-0.3627***	(0.1014)	-0.5071***	(0.1170)
		Non-teachers	-0.6878***	(0.0821)	-0.8743***	(0.0985)
	Age^2	Teachers	0.0055***	(0.0014)	0.0075***	(0.0016)
		Non-teachers	0.0091***	(0.0011)	0.0114***	(0.0013)
	AFQT	Teachers	0.0000	(0.0034)	-0.0180***	(0.0048)
		Non-teachers	-0.0342***	(0.0028)	-0.0382***	(0.0041)
	Schooling	Teachers	-1.3067***	(0.0690)	-1.6577***	(0.0771)
		Non-teachers	-0.8234***	(0.0491)	-1.0115***	(0.0547)
	Married dummy	Teachers	-0.4823**	(0.2315)	-0.0660	(0.3370)
		Non-teachers	-0.9525***	(0.1436)	-0.7105***	(0.2294)
	Married × male	Teachers	1.0247**	(0.5155)	1.3943**	(0.6537)
		Non-teachers	0.8969***	(0.3199)	0.5662	(0.3737)
	Children dummy	Teachers	-0.5902**	(0.2315)	-0.1275	(0.2758)
		Non-teachers	-0.7278***	(0.1436)	-0.6901***	(0.1841)
	Children × male	Teachers	0.6946	(0.4808)	0.3260	(0.5763)
		Non-teachers	-0.0254	(0.2981)	-0.4097	(0.3470)
	Occupation switching cost	Teachers	-2.2528***	(0.1621)	-2.2169***	(0.1756)
		Non-teachers	-2.2973***	(0.1722)	-2.2521***	(0.1816)
	Cost of returning to employment	Teachers	-4.6064***	(0.2328)	-4.5773***	(0.2821)
		Non-teachers	-4.4184***	(0.1126)	-4.3267***	(0.1231)

Table 5: Occupation specific utility function parameters by type

			One type		Seven types	
			Coefficient	Stand. Error	Coefficient	Stand. Error
Type 1	constant	Teachers	24.1067***	(1.9166)	28.9967***	(2.6967)
		Non-teachers	23.7474***	(1.5050)	31.6067***	(2.2058)
	$E[w]$		1.5810***	(0.1088)	0.9897***	(0.2626)
	$E[w]^2$		-0.0402***	(0.0041)	-0.0061	(0.0092)
Type 2	interaction	Teachers			1.6479	(2.1350)
		Non-teachers			-2.6042	(2.0928)
	$E[w]$				3.2084***	(0.5728)
	$E[w]^2$				-0.1363***	(0.0516)
Type 3	interaction	Teachers			6.4219***	(2.2981)
		Non-teachers			1.0854	(2.4902)
	$E[w]$				1.9310**	(0.9432)
	$E[w]^2$				0.0298	(0.1094)
Type 4	interaction	Teachers			4.5618**	(2.0874)
		Non-teachers			-2.4229	(2.3522)
	$E[w]$				0.8402***	(0.2367)
	$E[w]^2$				-0.0065	(0.0063)
Type 5	interaction	Teachers			-1.7169	(2.4202)
		Non-teachers			-8.8825***	(2.5508)
	$E[w]$				2.4700***	(0.3433)
	$E[w]^2$				-0.0618***	(0.0118)
Type 6	interaction	Teachers			-2.4901	(1.8972)
		Non-teachers			-7.4707***	(1.8423)
	$E[w]$				3.8278***	(0.3536)
	$E[w]^2$				-0.1642***	(0.0232)
Type 7	interaction	Teachers			-0.7483	(1.9378)
		Non-teachers			-6.1109***	(1.9234)
	$E[w]$				2.4667***	(0.3142)
	$E[w]^2$				-0.0574***	(0.0165)

Not surprisingly returns to schooling are higher for teachers. As previously discussed, teachers with more years of education receive a higher salary. These coefficients are very similar when we increase the number of types. Also, for both occupation groups, among single and childless individuals, males' earnings are higher. The gender earnings gap for teachers increases and decreases for non-teachers with seven types, and is similar for both occupations in this specification.

Married individuals in the non-teaching occupation have higher earnings than non-married individuals. This difference is not statistically significant for teachers. Not surprisingly, females with children have lower earnings in both occupations than females with no children. However, earnings for males with children are not different for teachers and are even higher for non-teachers compared to males with no children. These coefficients are qualitatively similar in both models presented.

A key to identification of the coefficient on earnings in the utility function is to have a variable which is only in the log earnings regression. We use sector-specific experience as the exclusion restriction. The coefficients on experience basically do not change when including more types. Not surprisingly, teaching experience is particularly relevant for teachers. Teacher earnings increase with teaching experience and continue to do so beyond the range of experience found in our data. Small levels of teaching experience provide little benefit outside of teaching. However, our point estimates suggest that teachers with considerable experience benefit in other jobs.[14] Overall, an extra year of teaching increases yearly earnings around \$1,600 for the average teacher, while it has an insignificant negative effect on yearly earnings for the average non-teacher.

Experience in other occupations increases earnings for both teachers and non-teachers at a decreasing rate but throughout the relevant experience range non-teaching experience is more valuable outside of teaching.

Finally, the standard deviations are modeled as linear functions of age (i.e., $\sigma = \sigma_A + \sigma_B \cdot age$). The estimates suggest that the variance of teachers' earnings is lower than the variance of non-teachers' earnings for all age groups.

We comment on earnings differences among the seven types in the next subsection.

[14] We expect that this reflects very experienced teachers transitioning to other well-paid jobs in education.

4.4 Characteristics of the types

Table 7 displays the average risk aversion and AFQT of the different types, the difference in log earnings equations of each type relative to type 1 in each occupation and the proportions of each type in the teaching and non-teaching groups. The rows are ordered by the size of the type effects in the teachers log earnings equation. Thus type 1 has, *ceteris paribus*, the highest earnings among teachers and type 3 the lowest. The estimates suggest that type 1 has an earnings advantage with respect to other types in the teaching occupation; her teacher earnings are higher than other type's teacher earnings all else equal. Type 1 is not particularly risk averse and has an AFQT slightly above average. For the non-teaching occupation, type 4 is the type with the greatest earnings advantage. Moreover, type 4 is the type with the highest average AFQT and lowest risk aversion.

Combining this information, we see that the distribution of types across occupations plays a modest role in the earnings gap between teachers and non-teachers. To see this we can calculate that the average teacher type earns 85 log points less than a type 1 in teaching compared with 19 log points for non-teachers relative to a type 1 in non-teaching. Most of the difference is due to how the types are rewarded. If the distribution of teachers by type were the distribution among non-teachers, their mean earnings would be roughly 13 percent higher. If the type distribution of non-teachers were that of teachers, their mean earnings would be 11 percent lower.

Moreover, there is more variation in earnings across types outside of teaching than within teaching. The weighted standard deviation of the type effect in teaching is about .28 while it is about .36 outside of teaching. Types that are good at non-teaching also tend to have high earnings in teaching. The correlation is about .87 using the distribution of teachers across types and .72 using the non-teachers' distribution. In short, there is a strong positive correlation between the (to us) unmeasured skills that raise earnings within and outside teaching, but these skills are rewarded more generously outside teaching.

In teaching the correlation between the type earnings coefficient and the average risk aversion of the type is close to zero, while in non-teaching there is a modest negative relation. More importantly, while we see almost no correlation between a type's average AFQT and

Table 6: Log earnings parameters

		One type		Seven types	
		Coefficient	Stand. Error	Coefficient	Stand. Error
AFQT	Teachers	-0.0005**	(0.0002)	0.0017***	(0.0003)
percentile	Non-teachers	0.0041***	(0.0001)	0.0034***	(0.0003)
Schooling	Teachers	0.1053***	(0.0046)	0.0878***	(0.0045)
	Non-teachers	0.0689***	(0.0014)	0.0633***	(0.0025)
Male dummy	Teachers	0.0332*	(0.0180)	0.0906***	(0.0295)
	Non-teachers	0.1590***	(0.0066)	0.0896***	(0.0140)
Married dummy	Teachers	0.0186	(0.0209)	-0.0223	(0.0225)
	Non-teachers	0.0692***	(0.0090)	0.0296***	(0.0109)
Married × male	Teachers	0.0665	(0.0543)	-0.0123	(0.0497)
	Non-teachers	0.0354***	(0.0118)	0.0493***	(0.0133)
Children dummy	Teachers	-0.1131***	(0.0215)	-0.1511***	(0.0176)
	Non-teachers	-0.0507***	(0.0089)	-0.0565***	(0.0098)
Children × male	Teachers	0.0857	(0.0570)	0.1186***	(0.0446)
	Non-teachers	0.1326***	(0.0112)	0.1469***	(0.0116)
Experience	Teachers	0.0372***	(0.0036)	0.0397***	(0.0028)
teaching	Non-teachers	-0.0031	(0.0024)	0.0066***	(0.0017)
Experience	Teachers	-0.0004***	(0.0002)	-0.0006***	(0.0001)
teaching^2	Non-teachers	0.0011***	(0.0002)	0.0006***	(0.0001)
Experience	Teachers	0.0266***	(0.0021)	0.0274***	(0.0020)
non-teaching	Non-teachers	0.0811***	(0.0018)	0.0760***	(0.0016)
Experience	Teachers	-0.0008***	(0.0001)	-0.0007***	(0.0001)
non-teaching^2	Non-teachers	-0.0016***	(0.0001)	-0.0015***	(0.0000)
sigmaA	Teachers	0.2272***	(0.0163)	0.3743***	(0.0146)
	Non-teachers	0.1948***	(0.0084)	0.2651***	(0.0052)
sigmaB	Teachers	0.0030***	(0.0004)	-0.0042***	(0.0004)
	Non-teachers	0.0083***	(0.0002)	0.0018***	(0.0001)
constant	Teachers	-0.0507***	(0.0089)	9.5617***	(0.0895)
	Non-teachers	8.7513***	(0.0255)	9.1146***	(0.0450)

Trend, trend squared and race dummies were also included.
Types interactions for the seven types model are shown in Table 7.

Table 7: Average characteristics, interaction coefficients and proportions per type

	Risk aversion standardized (1)	AFQT standardized (2)	Earnings Coefficients Teachers (3)	Earnings Coefficients Non-teachers (4)	Proportions Teachers (5)	Proportions Non-teachers (6)
Type 1	-0.0097	0.0424	-	-	0.0315	0.1559
Type 5	0.0055	-0.0780	-0.5434	0.1667	0.1164	0.1255
Type 4	-0.1991	0.3423	-0.6122	0.4728	0.0246	0.0963
Type 7	0.0321	-0.0409	-0.7580	-0.2044	0.3423	0.2692
Type 6	-0.0264	-0.1707	-0.9438	-0.4419	0.3025	0.2032
Type 2	0.0449	0.1133	-1.1725	-0.6926	0.1538	0.1095
Type 3	0.2599	0.1891	-1.6994	-0.9376	0.0289	0.0404

its earnings coefficient in teaching, there is a modest positive relation outside teaching. This explains why the differential return to AFQT is smaller with seven-types than in the homogeneous-type model.

5 Goodness of fit

It is common in structural papers to examine how well the model matches the data by displaying figures that compare predicted and observed averages. Some papers also calculate in-sample goodness-of-fit statistics. We show that, using these comparisons and statistics, all of our models seem to fit most data patterns well: occupation choices, earnings and characteristics of transitioning individuals. Then, we test the accuracy of out-of-sample predictions using the 80-20 cross-validation division explained in section 4.1. We randomly divide our sample into two groups, consisting of 80% and 20% of individuals. We re-estimate the models using the larger sub-samples. Then, we use the new coefficients and the data from the other 20% of individuals to calculate and test the predictions of our nine models. We show that even testing out-of-sample predictions, other than the log-likelihood discussed in section 4.1, no single model clearly outperforms the others.

5.1 In-sample fit

We begin by describing the goodness of fit in-sample. Figures 4 and 5 show, for the seven-types model, the percentage of teachers and non-teachers by age. Visually the model fits

well. These figures are very similar for all of the models estimated. Moreover, for every model the χ^2 goodness-of-fit statistic for each of the choices is below the 5% critical value, also suggesting that the models fit well. There are two important caveats: first it assumes that observations are independent. Yet, given strong persistence in choices, ignoring correlation over time tends to exaggerate the value of the χ^2 statistic.[15] Second this χ^2 statistic is not adjusted for the fact that it uses estimated parameters to calculate choice probabilities. Therefore it does not have a chi-squared limiting distribution (Moore, 1977).

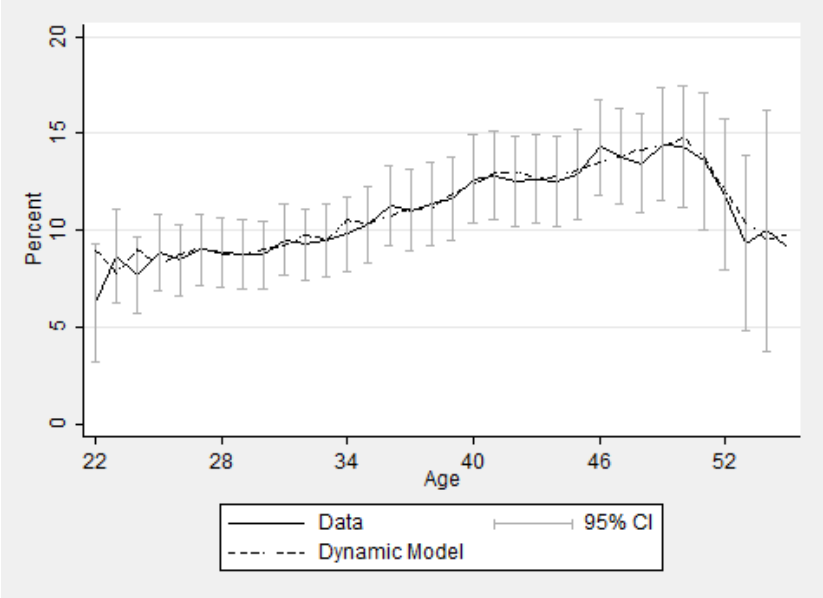


FIGURE 4: Percentage of teachers by age (seven-types model)

[15]Consider the extreme case where individuals were always teachers or non-teachers. We would effectively be exaggerating the number of observations by a factor equal to average number of years in the sample.

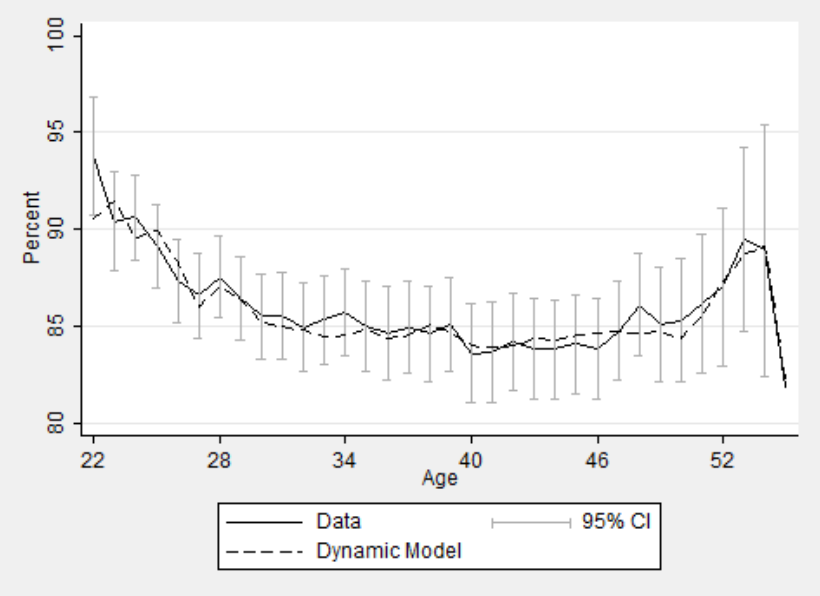


FIGURE 5: Percentage of non-teachers by age (seven-types model)

Table 8 shows average earnings for teachers and non-teachers. The model matches the average earnings of both teachers and non-teachers well in the sense that the predicted means lie within the confidence interval of the estimated population means from the data.[16] This is also true for the other models; the average earnings predicted lie within the confidence interval of the population means. Figure 6 depicts the log earnings by age for both groups, for the seven-types model. The absolute deviations from the data for different ages go from 0.0005 to 0.3984 for non-teachers and from 0.0005 to 0.3027 for teachers. These numbers are not very different for the other models.[17]

[16]Standard errors for the model will be calculated for next draft.

[17]The surprising ability to fit the sawtooth pattern at older ages reflects the fact that older workers are only observed in later surveys which were conducted only every other year. Thus the sample for 42 and 44 year olds is roughly constant and disjoint from the one for 43 and 45 year olds. The predictions in each case rely on the observations from in the relevant age group.

Table 8: Average earnings (in $1,000)

	Teachers	Non-teachers
Data	51.0941	76.6383
	(1.3468)	(1.4160)
Model	51.5034	75.4203

NOTE: Standard errors in parenthesis, clustered at the individual level.

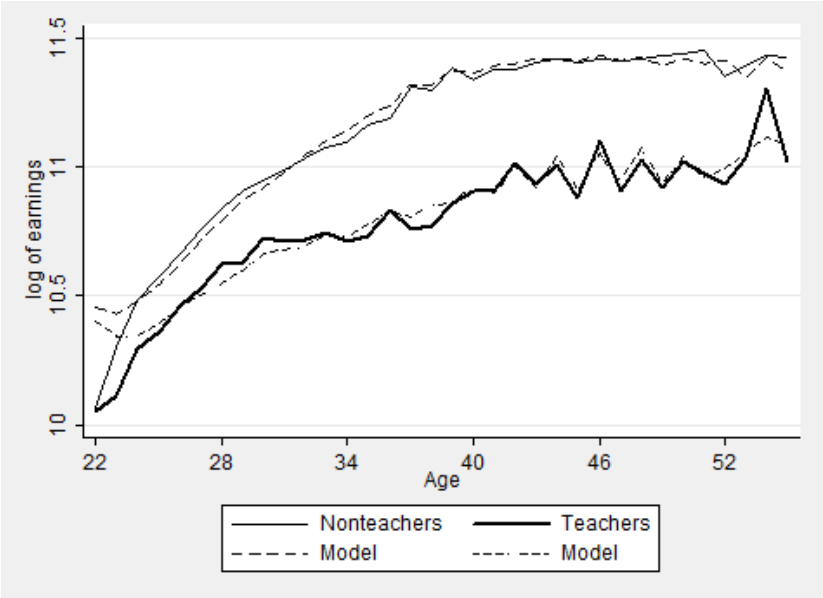


FIGURE 6: Log earnings by age (seven-types model)

Comparing tables 2 and 9, we see that the seven-types model replicates well transitions for non-teachers and not employed individuals. However, it over-predicts teachers' transitions into non-employment and under-predicts teachers' transition to non-teaching occupations. Subject to the caveats raised earlier, using a standard goodness of fit test, we reject the model's fit for transitions for teachers but not for non-teachers and individuals not working. The rejection is largely driven by our under-predicting transitions from teaching to non-teaching employment.

Finally, Table 10 shows the average characteristics of transitioning individuals. The model does a good job at matching observed averages in every cell. For every average, the model

Table 9: Transition matrix of predicted choices (percentages)

		predicted choice t		
		Teacher	Non-teacher	Not working
choice $t-1$	Teacher	93.8	3.9	2.3
	Non-teacher	0.7	98.4	0.9
	Not working	4.3	23.9	71.8

prediction is within the 95 percent confidence interval for the population average.[18] The same applies for the other models estimated. An interesting pattern, well replicated by the models, is that individuals who are teachers in one year and continue as teachers the following year are very risk averse and have low cognitive skills. As for non-teachers who do not change occupation, both the observed and predicted averages show that they are not particularly risk averse and have an AFQT score slightly above average.

5.2 Cross-Validation[19]

While, except for underestimating teachers transitions into other occupations, our models fit well, at least visually, within sample, a fairer test is their ability to predict out-of-sample in a cross-validation exercise. We take twenty random samples consisting of eighty percent of the individuals in our sample, reestimate the models and calculate how accurately we predict choices and log earnings for that part of the sample that was not used in the estimation.

When we predict occupation choice, for models with four types or more, we find that in twenty out of the twenty sub-samples the χ^2 goodness-of-fit statistic is less than the critical value at a 5% confidence level. Of course, the two caveats raised earlier apply; the χ^2 statistic does not consider the possible correlation of observations or the fact that we are using estimated parameters to calculate choice probabilities. In any case, this approach is not helpful in distinguishing among the models.

To assess how well our models predict earnings we follow the spirit of Table 8; we calculate average earnings for both groups, teachers and non-teachers, in every sub-sample and compare these averages with the out-of-sample predictions of our models. In the model with no

[18] Using standard errors clustered at the individual level to account for possible correlation of observations.
[19] This section is not yet complete for the nine-type model.

Table 10: Average standardized risk aversion and AFQT transitions

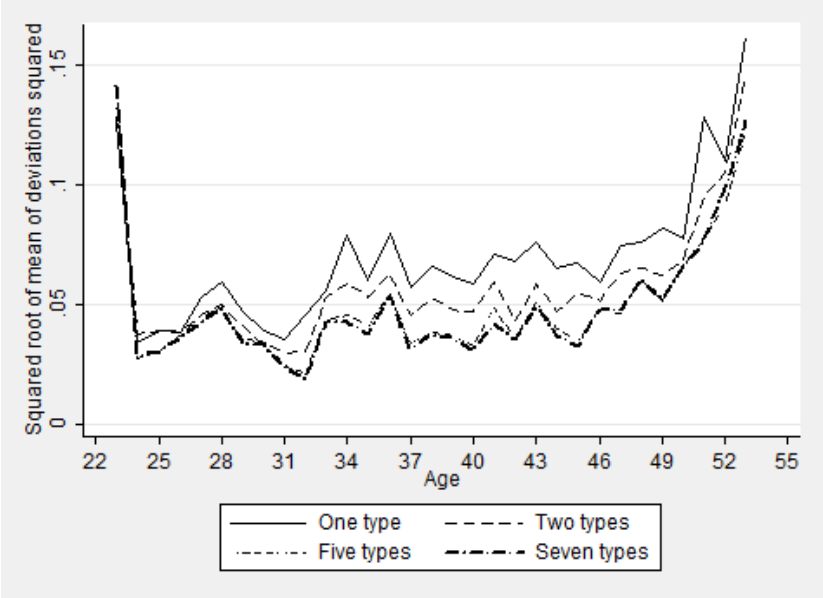
				t		
				Teacher	Non-teacher	Not working
$t-1$	Teacher	Risk aversion	Data	0.3183	0.1624	0.1417
			Model	0.3286	0.1729	0.2199
		AFQT	Data	-0.3078	-0.4177	-0.3878
			Model	-0.3119	-0.2717	-0.1338
	Non-teacher	Risk aversion	Data	0.1220	-0.0310	-0.1173
			Model	0.1703	-0.0305	-0.0542
		AFQT	Data	-0.3819	0.0413	0.0357
			Model	-0.2277	0.0382	0.0917
	Not working	Risk aversion	Data	0.2488	-0.2116	0.0715
			Model	0.2035	-0.0574	0.0512
		AFQT	Data	-0.3660	0.0618	0.0601
			Model	-0.2962	0.1266	0.0472

NOTE: * means the population average is statistically different from the point estimate using SE clustered at the individual level.

unobservable heterogeneity, in fourteen of the twenty sub-samples for at least one of the two occupations, we reject the hypothesis that the predicted average is within the confidence interval of the observed population average of that group. For the two and three types models, for seven of the twenty sub-samples we reject the hypothesis that predicted teachers' average earnings is within the confidence interval of the observed population average. For four types we reject this hypothesis in five of the twenty sub-samples, and for five types only in two of the twenty sub-samples. For every other model, with six, seven or eight types, we reject this hypothesis in one or none of the twenty sub-samples.

Additionally, we analyze the deviations of predicted from observed earnings by age for the twenty sub-samples. We summarize this information in Figure 7. These are unconventional graphs; they show the root mean squared error (square root of the mean of deviations squared for the twenty sub-samples) from the one, two, six and seven-types models. There is little difference between the six types and three, four, five, and eight-types models; so we do not present them. We also omit ages below 23 and above 53 years where deviations are larger due to a small number of observations. The "average deviations" depicted by age show a moderate decrease when comparing the one type model and any of the models with unobservable heterogeneity.

(a)



(b)

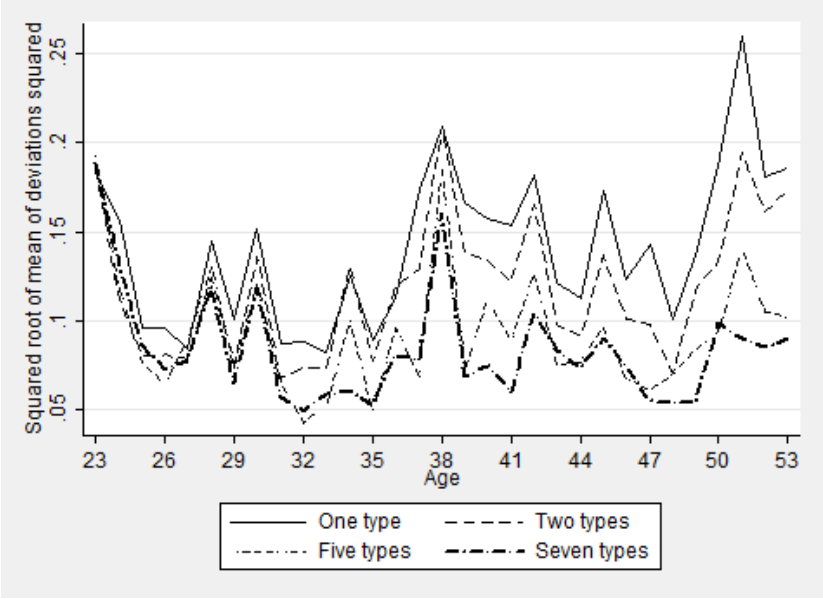


FIGURE 7: Root mean squared error

(a) Non-teachers

(b) Teachers

Finally, we apply our approach to predicting the characteristics of transitioning individuals. In each case, we ask whether the prediction from the model matches the population mean allowing for the confidence interval around the estimate of the population mean but not the imprecision of the model estimate. We calculate the t-statistic for the hypothesis that

Table 11: Maximum t-statistic of the 20 draws

			t		
			Teacher	Non-teacher	Not working
$t-1$	Teacher	Risk aversion	0.3502	3.5865**	2.9524*
		AFQT	0.6178	8.0887**	3.2988**
	Non-teacher	Risk aversion	3.3524**	0.3805	6.4889**
		AFQT	6.5224**	0.4223	6.2087**
	Not working	Risk aversion	3.8296**	4.0250**	1.8906
		AFQT	3.3307**	3.5502**	0.7637

NOTE: ** denotes cells where averages are statistically different, * denotes cells where averages could be statistically different.

the population average equals the model point estimate.[20] Table 11 shows the maximum t-statistic of the twenty draws for every cell for the seven-types model. Since we are performing multiple tests, we use Bonferroni's correction. If there is no correlation among the tests, the adjusted critical value at the 5% confidence level is 3.0233. Since the tests are calculated with sub-samples drawn from the same master sample there should be some correlation among tests, so we can use 3.0233 as an upper bound and 1.96 as a lower bound. That is, we definitely reject the hypothesis that model and data averages are the same for cells with a t-statistic higher than 3.0233, and we definitely cannot reject (at the .05 level) the hypothesis that model and data averages are statistically equal for cells with a t-statistic lower than 1.96. In the diagonal cells we definitely do not reject that predicted and observed averages are equivalent. On the other hand we can clearly reject the null for average AFQT and risk aversion of transitioning individuals. The results are almost identical for all models.

[20] Using standard errors clustered at the individual level.

6 Simulation

An advantage of structural modeling over reduced form estimation is the ability to perform counterfactual experiments. However, the credibility of such simulations relies on the model being a 'good enough' approximation of reality. When allowing for unobserved individual heterogeneity there are two risks. If we do not allow for sufficient heterogeneity, the model is misspecified. If we allow for excess heterogeneity, although the estimates remain consistent, our counterfactual estimates may suffer from overfitting the original model. Therefore, choosing the correct number of types is crucial, particularly if the results from simulations differ depending on the number of types.

In this section we show that the conclusions from our counterfactual experiment are highly sensitive to the number of types. Using our nine specifications we simulate how teachers' characteristics, in particular AFQT and risk aversion, would vary if a different contract were offered. To assess the experiment's 'cost' we also calculate average earnings given the new environment. We are interested in analyzing the change in teacher composition keeping quantity constant, that is we want to keep average probabilities of being a teacher and a non-teacher unchanged. To achieve this we adjust the constant terms in the log earnings equations for each simulation.

To be clear, we are not simulating an ideal policy. Such a policy would require objective and/or subjective measures of teaching performance that are not available in our data for teachers, let alone non-teachers. Instead we rely on the positive relation between teaching and other skills to justify this experiment.

Our simulation considers a world in which salaries for teachers reward observable and unobservable abilities in the same way as those of non-teachers. These changes imply an increase in the riskiness of the teaching occupation; there is much more variation in earnings across types outside of teaching than within teaching. To replicate this scenario we adjust several coefficients. First, we adjust teacher salaries so that they are as responsive to AFQT (our measure of "observable ability") and to schooling as those for non-teachers. Then, we replace the coefficients on the unobservable types in the log earnings equation for teachers with those from the non-teaching log earnings equation. Finally, we set the coefficient on

risk aversion in the teacher utility equation equal to its value in the utility equation for non-teaching. We also adjust the constant term so that the utility from teaching for individuals with the lowest risk aversion measure is unaffected by the increased riskiness.[21] We also raise teachers' earnings variance by setting the σ_A and σ_B parameters equal to the parameters estimated in the non-teaching earnings equation. Due to the conversion from logs to levels, this increases expected earnings in the selection equations.

In summary, we are changing: the risk aversion parameter in the teacher utility equation and the AFQT, schooling, variance coefficients and types coefficients in the teacher log earnings equation. Finally, we adjust the constant terms from log earnings equations so that there is only a change in the composition of teachers and not in the quantity (or average probability).

Table 12 shows the average effect of the simulation using the nine models. The results shown are the predicted averages of the simulation minus the predicted averages of each model. Although the simulated policy effects are imprecisely measured, the results differ dramatically depending on the number of types used, and the effects are not necessarily monotonic in the number of types. In the following paragraphs, we focus on a brief discussion of the point estimates and then we comment on whether the estimated differences are statistically significant from each other using the Hausman formula.

Average AFQT basically remains unchanged using the three types model and increases 0.36 standard deviations using the five types model. The model with no unobserved heterogeneity and the eight types models suggest the policy would increase average AFQT by 0.12 standard deviations.

Risk aversion, in turn, decreases by around 0.02 standard deviations with the one, two, and three types models and around 0.05 standard deviations with the seven, eight and nine types models. Yet, it decreases by more than 0.15 standard deviations with the five and six types models.

The changes in age, gender and race are negligible for the one through four types models and for the seven and eight types models. With five types, the change in average age is

[21] We make this adjustment because we would not anticipate that the expected utility of risk-neutral individuals to be affected by the increased riskiness. Given the phrasing of the question, we know only that this group has a low degree of risk aversion and not that its members are risk neutral.

nontrivial. It is particularly striking that the share of males increases dramatically in the simulation using the five and six types models. And we see a very notable decline in minority representation with five types and, to a lesser degree, with six types.

Interestingly, the differences in the simulation results when using the seven, eight and nine types models are rather small despite the large efficiency loss from moving to more than seven types.

To compare the predicted simulation effects of each model we use the Hausman formula and calculate the standard errors of differences between the predictions from the different models. The calculation is based on the argument that if the true model has N types, then the most efficient estimate of the policy effect is based on the model with this number of types. Models with fewer than N types are inconsistent while models with more than N types are consistent but not efficient. Consequently, under the null that there are N types, we can calculate the variance of the difference between the estimate using N types and, for example, N+1 types as the variance of the estimate using N+1 types minus the variance of the estimate using N types. And, we can test for equality of the parameters estimates from the two models using this formula.

As is common in such settings, we frequently find that the estimated standard error using the estimator that is consistent but not efficient under the null is, in fact, smaller than the standard error of the hypothesized efficient estimate. Such reversals provide informal evidence against the null hypothesis.

For four of the seven policy estimates in Table 12, the two-types model is more precise than the one-type model. For the remaining three, one of the three pairs of estimates is statistically significantly different. When we compare the three and two-types models, in two of the seven comparisons, the three-type model is more precise while the difference is statistically insignificant in the five remaining pairs. Similarly, in two of the seven comparisons, the four-types model is more precise than the three-types model and in one of the remaining five pairs, the difference is statistically significant.

However, when we compare five and four types, in one case the five-types estimate is more precise. In five of the six remaining cases, we reject equality of the estimates. Moreover, in six of seven cases, we reject equality of the estimates using five and two types. We view this

as strong evidence of preferring the estimates with five types to those with fewer types.

We also find no difference in the estimated effects using five or six types. But, when we compare seven with six types, the estimated policy effects in the former are more precise in six cases. Moreover, when we compare seven with five types, four of the estimated policy effects are more precise with the seven types model, and in two of the remaining cases the simulated effects are statistically significantly different between the models. Again, we view the evidence as suggesting that we require seven types.

Finally, we observe a large loss of precision when we move beyond seven types and in no case can we reject that the policy estimates are the same as those using the seven-type model. We view this as supporting our reliance on the seven-type model.

How much would such a change in policy cost? Here, again, the answer depends in important ways on the amount of heterogeneity we allow although the estimates are sufficiently imprecise that in no case can we reject the hypothesis that the policy estimates are equal, and in every case the confidence interval is large. The smallest 'cost' occurs when there are three types, in which case average teaching salaries increase by $1,700 per year or 3.3%. The estimated costs are also modest with one type (4.9%) and grow somewhat large with two (7.2%), four (8.4%), seven (5.2%), eight (7.2%) or nine types (6.1%). Our conclusions are strikingly different with five (18.1%) or six types (20.2%).

What accounts for these differences? Once we have more than one type, a large part of the change in the structure of earnings comes from the way that unobservable types are rewarded. As we include more types, the magnitude of the largest earnings gap for the type with the largest gap will tend to increase. This increases the value of switching from non-teaching to teaching for the group that benefits most from the change. Of course, this is not always the case, and the effect is partially offset by the tendency for the proportion of individuals in each type to fall. Thus the effect need not be monotonic. In our estimates and simulations with five and six types, we see a large shift of the most highly paid group into teaching.

Consistent with this explanation, the disruption required to effect such a policy is much greater in the simulations with five or six types. With one to three types we estimate that 12% - 13% of teachers (weighted by teaching years) would leave teaching. When juxtaposed with

an annual turnover rate of about 7%, this strikes us as manageable if the policy were phased in over an extended period. The level of turnover becomes somewhat more problematic (17% or higher) in the simulations with four or more types.

While the degree of disruption varies dramatically among the simulations, all suggest to varying degrees significant teacher resistance. In the homogeneous model, 36% of teachers (again weighted by years teaching) would be made worse off. With four types this grows to a majority, and with six types almost reaches 80%. Even though results are less extreme when we consider seven, eight or nine types, these three models also estimate the majority of teachers would be made worse off. Even in settings where teachers are not unionized, this would make a transition to this policy difficult.

Table 12: Average effect of simulations

	AFQT (standardized)	Risk av.	Age	Male	Black	Hispanic	Earnings (in $1,000)
DATA	-0.3067	0.2967	37.8	0.2578	0.1690	0.1382	51.1
	(0.0205)	(0.0175)	(0.2)	(0.0086)	(0.0074)	(0.0068)	(0.7)
One type	0.1162	-0.0233	0.0	0.0040	-0.0222	-0.0041	2.8
	(0.0207)	(0.0221)	(0.4)	(0.0107)	(0.0049)	(0.0021)	(8.0)
Two types	0.0410	-0.0176	-0.2	-0.0064	-0.0066	-0.0020	3.9
	(0.0250)	(0.0209)	(0.4)	(0.0131)	(0.0033)	(0.0015)	(10.1)
Three types	-0.0023	-0.0183	-0.3	-0.0047	-0.0026	0.0011	3.4
	(0.0434)	(0.0367)	(0.7)	(0.0267)	(0.0070)	(0.0012)	(10.0)
Four types	0.0023	-0.0391	-0.4	0.0012	-0.0031	-0.0040	4.3
	(0.0705)	(0.0538)	(1.0)	(0.0320)	(0.0047)	(0.0024)	(9.5)
Five types	0.3554	-0.2564	-2.5	0.2137	-0.0544	-0.0404	9.4
	(0.1284)	(0.0920)	(0.8)	(0.1125)	(0.0120)	(0.0182)	(14.4)
Six types	0.2335	-0.1489	-1.2	0.1298	-0.0406	-0.0186	10.4
	(0.2106)	(0.1853)	(1.4)	(0.2098)	(0.0218)	(0.0276)	(24.0)
Seven types	0.0746	-0.0396	-0.5	0.0063	-0.0145	-0.0046	2.5
	(0.1740)	(0.1122)	(1.8)	(0.0666)	(0.0096)	(0.0128)	(13.1)
Eight types	0.1187	-0.0586	-0.7	0.0340	-0.0181	-0.0070	3.8
	(0.2718)	(0.2033)	(3.6)	(0.1571)	(0.0102)	(0.0206)	(19.5)
Nine types	0.1048	-0.0595	-1.1	-0.0005	-0.0132	-0.0056	3.1
	(0.4257)	(0.2594)	(5.1)	(0.1868)	(0.0160)	(0.0429)	(24.5)

7 Conclusion

This paper contributes both to our substantive understanding of reforming teacher compensation and to the practice of structural modeling.

With respect to the latter, it is widely recognized that the strength of structural modeling is that it allows us to consider experiments that lie outside the data. It is equally widely recognized that the validity of the experiment relies on the model being (at least approximately) correct. We show that establishing that a model fits well within sample is, at best, weak evidence that it is approximately correct. Even showing by way of cross-validation that it fits well out-of-sample should not be convincing. In our case, even a model with no unobserved heterogeneity appears to fit well within sample and does reasonably well in cross-validation. It would be easy to conclude that allowing for two or three types is adequate to fit the data. But more formal methods do not support this conclusion. The Bayesian Information Criterion favors seven types. When we rank models in terms of their performance in cross-validation, we also conclude that we need many types. Even though we cannot reject that eight types is better than seven or that nine types is better than eight, in an informal sense, the differences between the two pairs of models are not large. As we show, the conclusions we draw from our experiment can depend crucially on choosing among models, all of which appear to fit well both within and out-of-sample. Interestingly, the conclusions drawn with seven or more types (i.e., when approaching what we believe is the requisite number of types) are almost identical. Moreover, we can always reject the equality of the policy estimates for five versus fewer types and then for seven versus fives types for at least some outcomes.

With respect to reforming teacher compensation, we establish that, among college graduates, teachers are not only drawn disproportionately from the lower part of the AFQT distribution, but they are also more risk-averse than their counterparts outside teaching. When we allow for unobserved heterogeneity, the low mean AFQT score among teachers reflects not a low return to cognitive skill within teaching but low returns to other skills, correlated with AFQT. The compression of earnings within teaching attracts relatively risk-averse individuals.

We show that if it were possible to revise compensation in teaching to mimic the return to

skills and riskiness of the non-teaching sector, there would be a modest increase in average teachers' AFQT and a modest decrease in average risk aversion. However, such a shift would adversely affect many of those who are currently in teaching and who would suffer large utility losses if they shifted out of teaching. This makes the process of reform challenging.

References

Aguirregabiria, Victor and Pedro Mira, "Dynamic discrete choice structural models: A survey," *Journal of Econometrics*, 2010, *156* (1), 38–67.

Arcidiacono, Peter, "Ability sorting and the returns to college major," *Journal of Econometrics*, 2004, *121* (1), 343–375.

_ **and Paul B Ellickson**, "Practical methods for estimation of dynamic discrete choice models," *Annu. Rev. Econ.*, 2011, *3* (1), 363–394.

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

Biasi, Barbara, "Unions, salaries, and the market for teachers: Evidence from Wisconsin," 2017.

Dolton, Peter J, "Teacher supply," *Handbook of the Economics of Education*, 2006, *2*, 1079–1161.

Eckstein, Zvi and Kenneth I Wolpin, "Why youths drop out of high school: The impact of preferences, opportunities, and abilities," *Econometrica*, 1999, *67* (6), 1295–1339.

Hansen, Peter Reinhard and Elena-Ivona Dumitrescu, "Parameter estimation with out-of-sample objective," 2016.

Hanushek, Eric A, John F Kain, and Steven G Rivkin, "Why public schools lose teachers," *Journal of human resources*, 2004, *39* (2), 326–354.

Hout, Michael and Editors Elliott Stuart, *Incentives and test-based accountability in education*, National Academies Press, 2011.

Hoxby, Caroline M and Andrew Leigh, "Pulled away or pushed out? Explaining the decline of teacher aptitude in the United States," *The American Economic Review*, 2004, *94* (2), 236–240.

Keane, Michael P and Kenneth I Wolpin, "The career decisions of young men," *Journal of political Economy*, 1997, *105* (3), 473–522.

Lazear, Edward P., "Performance pay and productivity," *American Economic Review*, 2000, *90*, 1346–1361.

Leigh, Andrew, "Teacher pay and teacher aptitude," *Economics of Education Review*, 2012, *31* (3), 41–53.

Moore, David S, "Generalized inverses, Wald's method, and the construction of chi-squared tests of fit," *Journal of the American Statistical Association*, 1977, *72* (357), 131–137.

Rust, John, "Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher," *Econometrica: Journal of the Econometric Society*, 1987, pp. 999–1033.

Stinebrickner, Todd R, "Compensation policies and teacher decisions," *International Economic Review*, 2001, *42* (3), 751–780.

_ , "A dynamic model of teacher labor supply," *Journal of Labor Economics*, 2001, *19* (1), 196–230.